Statistical learning methods for functional data with applications to prediction, classification and outlier detection

by

Nicolás J. Hernández Banadik

A dissertation submitted by in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Business and Quantitative Methods

Universidad Carlos III de Madrid

Advisor:
Dr. Alberto Muñoz García

July, 2019

A mi abuelo, el fabricante de juguetes.

Acknowledgements

I would like to thank my friends and family, whose support encouraged my to pursue this objective. To my PhD. advisor and friend Prof. Dr. Alberto Muñoz for sharing with me all his knowledge. Also I would like to thank Dr. Gabriel Martos for being an inspiring role model as a young researcher to follow.

Prof. Dr. Julien Jacques and Dr. Jairo Cugliari my advisors in Lyon for the time and effort dedicated to my work. I also want to acknowledge the financial support received from the Spanish Ministry of Economy and Competitiveness ECO2015-66593-P and the UC3M PIF scholarship for doctoral studies.

Agradecimientos

Los agradecimientos son la parte más divertida de escribir. Cuando uno lee los agradecimientos de una tesis, están: los que enumeran una lista de nombres, y los otros. Ya pueden hacer la primer inferencia del día. Armamos rampas y cross–check.

Desde que empecé este proceso allá por el 2013 hasta el día de hoy, siento que me han abierto el cráneo y me han dado con un hierro en el cerebro en muchos aspectos. Uno puede pensar que han sido los más de cien artículos y los pocos libros que me he leido. Yo creo que sin duda fue gracias a mi director y amigo Alberto. Teñidas de un humor digno de *La hora chanante*, nuestras a veces un poco tensas reuniones de trabajo, y la dedicación de Alberto, han sido el motor que me ha permitido rodar hasta el V1, rotate.

La meta emigración se le puede llamar al proceso de emigrar dentro de una emigración ya realizada anteriormente. Tuve una pisca de esto cuando estuve en Lyon, Francia. Por momentos desde la soledad de la colina de Fourvière, y por otros desde el horno al que tenian por "oficina de investigadores visitantes" en el laboratorio ERIC; le di un empujón fuerte a la tesis. Allí tuve el gusto de trabajar y recibir consejo de Julien Jacques quien ha sido de enorme ayuda en mi proceso de buscar trabajo. También compartí con un compatriota de nombre divertido y gran sentido del humor, Jairo Cugliari. Jairo me desafió día a día de la estancia a estar al límite de mis capacidades. De la hermosa ciudad de Lyon, me llevé muy buenos recuerdos de Antoine, un garca como pocos, y de Margot, que aunque de agenda apretada, supo compartir unas cuantas pintas.

Jeje, me estoy acordando cuando Gabi se moría de la de risa con mis comentarios estúpidos alguna que otra vez que pateamos las calles de Malasaña. A él le debo su dedicación y obstinación porque aprendiera algo. Siempre con gran sentido de la paciencia, aún cuando yo mismo me hubiera mandado a la esquina a ver si llueve. Es como mi segundo director. Un verdadero Crack, *chapeau*.

Hacer una tesis no es solo escribir un conjunto de simbolos griegos que algunos pocos entienden y a muchos menos les interesa. También es un proceso –el cual quise abortar varias veces en pos de dedicarme a ser guia turístico por el Madrid de los Austrias– que requiere de cierta inteligencia emocional. No aborté. La inteligencia emocional que he adquirido se la debo a mi compañera y amante, Hanna. Con ella he compartido noches de excesos, risas, llantos, viajes y los más hermosos recuerdos. Una luchadora insaciable que me ha enseñado mucho y espero que lo siga haciendo. Se la acusa de varios intentos de pegarme una patada en el culo y algunos intentos de fuga, pero estos delitos prescriben pronto. También se la acusa crímenes de lesa humanidad como comprenderme, apoyarme, escucharme, escucharme, escucharme, amarme incondicionalmente y otros menesteres que no vienen al caso.

Mi familia ha sido secuestrada emocionalmente. Antes de agradecerles, quiero pedirles perdón. Perdón por todas esas despedidas en el aeropuerto de Montevideo que de sólo recordarlas me hacen un nudo en la garganta. Gracias a los Hernández y a los Banadik –forma curiosa de agradecer a una familia tan chica–. Gracias viejo por el apoyo de siempre, por no dejarme bajar los brazos nunca. Gracias mami por bancártela como una generala y perdón por no poder responder nunca a tu tan repetida pregunta de "¿cuándo vas a volver?". Gracias Manu, mi cable a tierra; el tipo que hace todo bien. Te admiro mucho hermano. Otro verdadero Crack, *chapeau*. Gracias abuelo por tu fortaleza y esa historia de vida inspiradora.

Porque no solo de artículos y códigos vive el doctorando, también están las cerverzas y la fiesta. La birra y la joda van de la mano y siempre en compañía de amigos. También conocido como *el cuarteto del disco feliz*, junto con el Emo, el Salo y el Vara nos hemos deshidratado de risa. Incondicionales amigos de hoy y siempre. En mis regresos a Uruguay se hacen un lugar para compartir momentos inolvidables. Un abrazo para ustedes gurises, vamo' arriba! El Emo, un hermano que me regaló la vida. Despúes de unos cuantos amagues me vino a visitar. Le tocó lo peor y lo mejor. Lo adoro. No se cuántos cientos de kilómetros caminé mientras escuchaba los audios de Whatsapp de Fede. Su manera de estar presente es a través del buzón de pensamientos. Desde profundas conversaciones que harían reir Nietzsche, hasta las recetas más sabrosas, pasando por improperios a políticos y futbolistas; todo eso y más puede ud. encontrar en el buzón de pensamientos. Jose, el único joven investigador que ya tiene dos casas de su propiedad. Increíble. Amigo de toda la vida. Gran viajante que me ha hecho reír con sus anécdotas y alegrado con sus visitas.

Kandinksy, uno de mis pintores favoritos. El que está y no está. Una persona con admirable capacidad reflexiva, capaz de sembrar la duda en el más fiel de los fieles. Con el Guille hicimos 600 km en bici hasta Santiago de Compostela. Compañero de grandes "*abulonadas*" en las terrazas de Lavapiés –en doble jornada–. Con él tengo el record de cervezas en una tarde: 38 tercios. El Facu, un atleta. También compañero de pedaleadas por el norte y de cafés asquerosos. Por los patios de la universidad, junto con el Manu Ceballos haciamos el trío de las anécdotas nunca contadas, flatulencias variadas, relexiones sin sentido y bueno, también apreciábamos siluetas. Cuando compartíamos solecito con Luciana y Silvina, las conversaciones se volvían más sensatas pero no menos divertidas. Hermosas compañías.

En un momento decidí abandonar uno de los círculos del infierno de Dante y me mudé de Getafe a Madrid. Pasé a formar parte de la gloriosa *Embajada*, allí ha transcurrido mi vida junto a Sol, la Michelle Pfeiffer del Batman de Tim Burton. Wave, el ansioso menos ansioso del planeta, Chema el más cívico de los osos, y el Teniente. El teniente..., el hermano mayor que nunca tuve, un italiano en toda regla, consejero personal, asesor fiscal, eludidor nato de lavadas de plato, cocinero de salsas varias y gran relator de anécdotas. A Nacho, que inmoratlizó la frase "arriba el Sábado...", Gioia, Christina, todos los embajadores y agregados culturales de la Embajada. Salú!

No puedo terminar sin agradecer a la barra de los cachivaches. A todos los compañeros del doctorado y de oficina. Hoang, un chino que cuando llegó no sabía abrir el R, hoy es un crack y siempre dispuesto a ayudar; Jorge, Antonio, Ángela, Javi, Alba, Juanmi, Paco y Susan. A Fanny y Andrea por los domingos de sofá. En fin, grazie a tutti. Ascenso positivo, tren arriba, y a volar!

–

PD: a lo largo de este texto habrán encontrado algunos *gags* relacionados con la aviación. Eso es por mi abuelo que le hubiera gustado que fuera piloto y terminé siendo...

Published and submitted contents

- Published contents:
  - Hernández N., Muñoz A. (2016). Kernel Depth Measures for Functional Data with Application to Outlier Detection. In: Villa A., Masulli P., Pons Rivero A. (eds) Artificial Neural Networks and Machine Learning – ICANN 2016. ICANN 2016. Lecture Notes in Computer Science, vol 9887. Springer, Cham
    - Co–author.
    - It is partially included in chapter 3 of the thesis.
    - The material from this source included in the thesis is not indicated by typographical means or references.
  - Hernández N., Muñoz A. (2016). Kernel depth funcions for functional data. Universidad Carlos III de Madrid. Departamento de Estadística.
    - `http://hdl.handle.net/10016/24615`
    - Co–author.
    - It is partially included in chapter 3 of the thesis.
    - The material from this source included in the thesis is not indicated by typographical means or references.
  - Martos, G.; Hernández, N.; Muñoz, A.; Moguerza, J.M. (2018). Entropy Measures for Stochastic Processes with Applications in Functional Anomaly Detection. Entropy 2018, 20, 33.
    - `https://www.mdpi.com/1099-4300/20/1/33`
    - Co–author.
    - It is fully included in chapter 4 of the thesis.
    - The material from this source included in the thesis is not indicated by typographical means or references.
- Contents submitted for publication:

- Hernández N., Muñoz A and Martos, G. (2019). Kernel depth measures for functional data with applications in functional outlier detection. Patter Recognition.
  - Co–author.
  - It is partially included in chapter 3 of the thesis.
  - The material from this source included in the thesis is not indicated by typographical means or references.
- Hernández N., Martos, G. and Muñoz A (2019). Boosting classification performance with functional time series: A domain selection approach. Journal of Automatica Sinica.
  - Co–author.
  - It is fully included in chapter 6 of the thesis.
  - The material from this source included in the thesis is not indicated by typographical means or references.

Abstract

In the era of *big data*, Functional Data Analysis has become increasingly important insofar as it constitutes a powerful tool to tackle inference problems in statistics. In particular in this thesis we have proposed several methods aimed to solve problems of prediction of time series, classification and outlier detection from a functional approach.

The thesis is organized as follows: In Chapter 1 we introduce the concept of functional data and state the overview of the thesis. In Chapter 2 of this work we present the theoretical framework used to we develop the proposed methodologies.

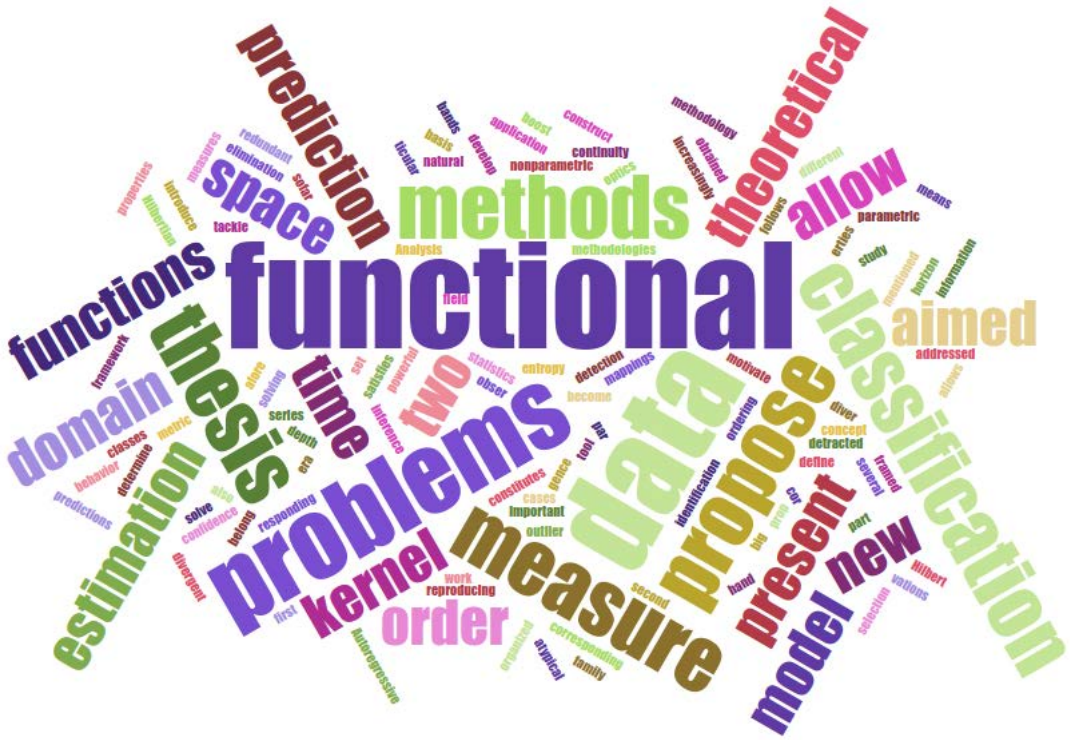
In Chapters 3 and 4 two new ordering mappings for functional data are proposed. The first is a Kernel depth measure, which satisfies the corresponding theoretical properties, while the second is an entropy measure. In both cases we propose a parametric and non-parametric estimation method that allow us to define an order in the data set at hand. A natural application of these measures is the identification of atypical observations (functions).

In Chapter 5 we study the Functional Autoregressive Hilbertian model. We also propose a new family of basis functions for the estimation and prediction of the aforementioned model, which belong to a reproducing kernel Hilbert space. The properties of continuity obtained in this space allow us to construct confidence bands for the corresponding predictions in a detracted time horizon.

In order to boost different classification methods, in Chapter 6 we propose a divergence measure for functional data. This metric allows us to determine in which part of the domain two classes of functional present divergent behavior. This methodology is framed in the field of domain selection, and it is aimed to solve classification problems by means of the elimination of redundant information.

Finally in Chapter 7 the general conclusions of this work and the future research lines are presented.

x


functional
prediction
methods
theoretical
allow
space
functions
thesis
classification
aimed
domain
time
two
problems
data
propose
estimation
kernel
measure
present
model
new
order

Resumen

En la era del *big data*, el Analisis Funcional de Datos ha cobrado cada vez mas relevancia en la medida en que constituye una herramienta muy potente a la hora de resolver problemas de inferencia estadística. En particular en esta tesis hemos propuesto diversas metodologías orientadas a solucionar problemas de predicción de series de tiempo, classficación y detección de atípicos desde un enfoque funcional.

El trabajo se encuentra organizado de la siguiente manera: En el Capítulo 1 introducimos el concepto de datos funcionales y motivamos los problemas abordados en la tesis. En el Capítulo 2 de este trabajo presentamos el marco teórico sobre el cual desarrollamos las metodologías propuestas.

En el Capítulo 3 y 4 se proponen dos nuevas medias de orden para datos funcionales. La primera de ellas es una medida de profunidad la cual satisface las propiedades teóricas correspondientes, mientras que la segunda es una medida de entropía. En ambos casos proponemos métodos de estimación tanto paramétricos como no paramétricos que nos permiten establecer un orden en los datos bajo estudio. Una aplicación natural de este tipo de medidas es la detección de observacioes (funciones) atípicas.

En el Capítulo 5 estudiamos el modelo autoregressivo funcional o autoregresivo Hilbersiano. Asimismo proponemos una nueva familia de funciones base para la estimación y predicción del modelo anteriormente mencionado, que pertenencen a un espacio de Hilbert con nucleo reproductor. Las propiedades de continuidad obtenidas en este espacio nos permiten la construcción de bandas de confianza para las correspondientes predicciones en un horizonte temporal detrminado.

Con el objetivo de potenciar distintos métodos de classificación para datos funcionales, en el Capítulo 6 proponemos una medida de divergencia para éste tipo de objetos. Esta métrica nos permite determinar en qué parte del dominio dos clases de datos funcionales tiene un comportamiento divergente. Esta metodología permite re-

solver problemas de selección de dominio, eliminando información redundante, y por tanto mejorar los resultados de clasicación.

Finalmente en el Capítulo 7 se presentan las conclusiones generales de este trabajo y las futuras líneas de investigación.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Functional Data Analysis (FDA) deals with the theoretical and methodological analysis of objects that can be expressed in the form of functions or more complex objects such as images. The nature of the object of analysis, in this case functions, is intrisically infinite–dimensional and is where relies the richness of the object under study.

Functional data can be defined as a set of random sample of independent real–valued elements on a compact interval $T = [a, b]$ –that can be assumed to be $[0, 1]$ withouth loss of generality–. This random sample is constituted by realizations of a stochastic process $X(t) \in L^2$, where $\mathbb{E} \int_T X^2(t)dt < \infty$.

A more formal definition can be stated by considering $(\Omega, \mathcal{F}, P)$ as the probability space where the random functions of interest are defined, where $\mathcal{F}$ is the σ-algebra in Ω and P a σ-finite measure. We consider random elements (functions) $X(\omega, t) : \Omega \times T \to \mathbb{R}$ in a metric space (T, τ). As usual in the case of functional data, the realizations of the random elements $X(\omega, \cdot)$ are assumed in $C(T)$, the space of real continuous functions in a compact domain $T \subset \mathbb{R}^d$ endowed with the uniform metric.

The process $X(\omega, t)$ is unobservable. Therefore in practice analysts and researchers need data to infer the charactesitics of this underlying process. In what follows, we will consider functional data as the set of discrete paths that constitutes realizations of the stochastic process $X(\omega, t)$ indexed in the closed subset T; also known in the literature as raw functional data Hsing and Eubank (2015). In that sense, raw functional data consist of the collection of the functions recorded over a fixed or random grid of discretized points, say $x(t_1), \ldots, x(t_m)$. The grid of discretized points usually is equally spaced – the elements differ by the same space between each other $t_i - t_{i-1} = t_{i+1} - t_i$–, which

asymptotically tends to zero as $m \to \infty$. Thus, analysis has to be conducted departing from the discrete version of the curve, $x(t_1), \ldots, x(t_m)$.

The terminology functional data was first adopted by Ramsay (1982) and extended for several author as Hall and Heyde (2014); Ramsay (2006); Ferraty and Vieu (2006); Hsing and Eubank (2015); Horváth and Kokoszka (2012) and references there in. A extended review is presented in Wang et al. (2016) and Cuevas (2014). Functional data analysis is of interest in fields such as genetics, chemometrics, physical processes, growth analysis, finance, among others.

To bring a visual inspection of the kind of data that FDA deals with, in Figure 1.1 are presented two well known data bases, data from the *Berkeley Growth Study*, and the *Canadian Weather* data. In the right panel of Figure 1.1 we observe the height in centimeters of 54 girls recorded at a common discretized set of 31 non–equidistant ages, between 1 and 18 years. In the right panel are illustrated the average daily temperatures in ºC of 35 weather stations located in different location across Canada, see Ramsay (2006) for further details.

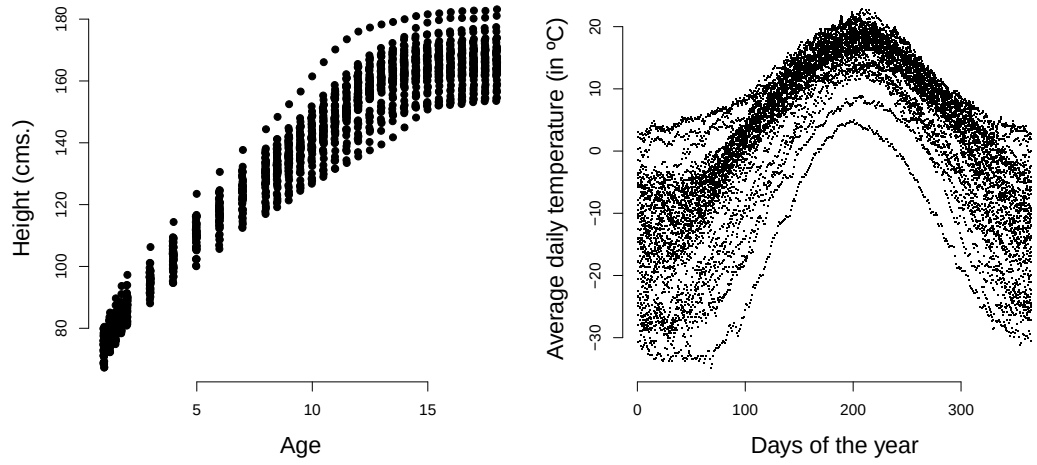


Figure 1.1: Growth Berkeley Study data (left). Canadian Weather data (right).

1.1 Overview of the thesis and contributions

Overview of Chapter 2

In Chapter 2 we present and discuss the concept of functional data and the implications on statistical inference. In particular we state the theoretical framework that we consider along the whole manuscript. All the statistical learning methods are based on a reproducing kernel Hilbert space model for functional data.

Contributions of Chapter 3

In the last years the concept of data depth has been increasingly used in Statistics and related fields as a center-outward ordering metric for multivariate and functional data sets. Many of the functional measures operate directly on the raw representation of the data which somehow contradicts its functional nature and also presents some weaknesses. Chapter 3 of this thesis introduce introduce kernel depth measures for functional data, i.e. realizations of a stochastic process, represented in a Reproducing Kernel Hilbert Space. Through this representation, complex objects such as time series and images are transformed into points in finite dimensional Euclidean spaces.

Two depth measures that induce order into the data were proposed: i) the Kernel Mahalanbobis Depth (KMD), based on the Mahalanobis distance jointly with a robustified version of it (RML–KMD) and the Generalized Kernel Depth (GKD) based on a generalization of the Mahalanobis depth via density kernels. This measure is valid for univariate and multivariate functional data, i.e. time series and images respectively. We prove that the proposed measure fulfils several desirable theoretical properties. Simulations results demonstrate that GKD works considerably better than other depth measures when the goal is to identify anomalous or outlier observations in non-Gaussian scenarios. Additionally we conduct several analyses of mortality rate curves, images processing, and biometric data as interesting applications of functional outlier detection.

Contributions of Chapter 4

Chapter 4 propose a definition of Entropy for stochastic processes, considering a Reproducing Kernel Hilbert Space model to estimate the Entropy from a random sample of realizations of a stochastic process, namely functional data, and introduce two approaches to estimate minimum entropy sets for functional anomaly detection. These

sets are relevant to detect anomalous or outlier functional data. A numerical experiment illustrates the performance of the proposed method; in addition, we conduct an analysis of mortality rate curves as an interesting application in a real-data context to explore functional anomaly detection. We also show the convergence of the parametric Entropy estimation method to the true values through a Monte–Carlo simulation. Morover the order invariance property is studied for both the parametric and non–parametric approach.

Contributions of Chapter 5

In Chapter 5 we present a new autoregresive Hilbertian model for functional time series. Based on a reproducing kernel Hilbert space framework, the first contribution is to develop a new family of basis functions to estimate the autocorrelation operator Ψ and to predict an entire new function for the whole domain. Throughout several Monte–Carlo studies, we show the performance of the proposed model, in terms of the root mean squared error, against well known prediction methodologies for functional time series.

As a second contribution, we tackle the issue of constructing predictive conifidence bands for the point forecast. We present a discussion related to the pointwise and simultaneous inference approaches to construct the predictive bands. Our proposed methodology is based on a model–based bootstrap approach for functional time series, which is an extension of the PRR Pascual et al. (2004) bootstrap procedure. We theoretically justify our proposal based on the continuity of the integral operator, noticing the advantage of the reproducing kernel Hilbert space framework over other approaches.

Contributions of Chapter 6

Domain Selection is embedded into the field of feature selection and implies the selection of the best functional data subinterval. Functional data and, in particular, time series, need to be transformed to finite dimensional data to apply standard statistical inference techniques.

In Chapter 6 we propose a novel *domain selection* methodology for functional time series data. We extend the concept of Kullback–Leibler divergence to drop out redundant information in time series and then select the best subinterval for classification

purposes. Here we consider a particular functional data analysis technique to obtain such a representation and then we use it on the subintervals obtained by Domain Selection to provide finite dimensional representations of the time series. In particular we introduce the *divergence curve* –and related concepts– as a tool to drop out redundant information in the context of supervised classification problems. Based on the quantiles of the empirical distribution of the *divergence curve* the proposed method learn and infer about the sub–interval of the domain that better discriminates the classes of functions. Simulations results show that the proposed methodology improves the classification performance reducing, at the same time, the computational burden of several functional classification methods. We apply the analysis to several functional time series data sets and the empirical results show remarkable improvements in supervised classification when the effective domain is learned in a first round of the problem.

Chapter 7

In Chapter 7 we summarize the work done in the thesis and its main contributions. We also point out the most important future research lines.

Chapter 2

Functional data framework

For the sequel let $(\Omega, \mathcal{F}, P)$ be a probability space, where $\mathcal{F}$ is the σ-algebra in Ω and P a σ-finite measure. We consider random elements (functions) $X(\omega, t) : \Omega \times T \to \mathbb{R}$ in a metric space (T, τ). As usual in the case of functional data, the realizations of the random elements $X(\omega, \cdot)$ are assumed in $C(T)$, the space of real continuous functions in a compact domain $T \subset \mathbb{R}$ endowed with the uniform metric.

The first and second moments of the stochastic process $X(\omega, t)$ defined in $C(T) \subset L^2(T)$ are defined by the mean function $\mu(t) = \mathbb{E}(X(\cdot, t))$, and the covariance operator $\Sigma(s, t) = cov(X(\cdot, s), X(\cdot, t))$. When the sample design is common for all the n observations the mean function and the covariance operator can be estimated by the sample versions: $\hat{\mu}(t_i) = \frac{1}{n}\sum_{j=1}^{n} x_j(t_i)$ and $\widehat{\Sigma}(t_k, t_l) = \frac{1}{n}\sum_{j=1}^{n}(x_j(t_k) - \mu(t_k))(x_j(t_l) - \mu(t_l))$, for $k \neq l$ and $i = 1, \ldots, m$.

A suitable representation for the stochastic process can be obtained by the expansion of the random paths $x_j(t)$ in a functional basis constituted by the eigenfunctions of the covariance operator $\Sigma(s, t)$. The first ones in getting this result were Karhunen (1946) and Loève (1946). Following the Karhunen–Loève expansion –see also (Bosq, 2012, pp. 25, Theorem 1.5)–, let $X(\omega, t)$ be a centered (zero–mean) stochastic process with continuous covariance function $K_X(s, t) = \mathbb{E}(X(\omega, s)X(\omega, t))$, then there exist a basis $\{e_i\}_{i\geq 1}$ of $C(T)$ such that for all $t \in T$

$$X(\omega, t) = \sum_{i=1}^{\infty} \xi_i(\omega) e_i(t), \tag{2.1}$$

where the sequence of random coefficients $\xi_i(\omega) = \int_T X(\omega, t)e_i(t)dt$ are zero mean random variables with (co)variance $\mathbb{E}(\xi_i \xi_j) = \delta_{ij}\lambda_j$, being δ_{ij} the Kronecker delta and

$\{\lambda\}_{j\geq 1}$ the sequence of eigenvalues associated to the eigenfunctions of $K_X(s,t)$. The equality in Equation 2.2 must be understood in the mean square sense, that is

$$\lim_{d\to\infty} \mathbb{E}[(X(\omega,t) - \sum_{i=1}^{d} \xi_i(\omega)e_i(t))^2] = 0, \tag{2.2}$$

uniformly in T. Therefore, we can always consider a ε–near representation $X_d(\omega,t) = \sum_{i=1}^{d} \xi_i(\omega)e_i(t)$ such that for all ε arbitrarily small, there exists an integer D such that for $d \geq D$ then $\tau(X, X_d) = \sup_{t\in T} |X(\omega,t) - X_d(\omega,t)| \leq \varepsilon$. From here, it is possible to establish a suitable way to elaborate inferential techniques for curves and any other functional data object that can be represented in this way since the goal is to model a d-dimensional vector of random variables: $\{\xi_i(\omega)\}_i^d$ obtained from $X_d(\omega,t)$.

The Karhunen–Loève expansion is known in the literature of functional data as the Functional Principal Component expansion. Nevertheless, the representation of functional data can be achieved by expanding the functions into other function bases such as splines, Fourier or Wavelets, see Wahba (1990); Ramsay (2006); Ferraty and Vieu (2006). In the next section we detail an alternative representation of functional data: The RKHS.

2.1 A reproducing kernel Hilbert space model for functional data

Following the usual approach in Functional Data Analysis Ramsay (2006); Ferraty and Vieu (2006), to represent curves which are infinite-dimensional objects by nature, we must choose an orthonormal bases of functions $B = \{\phi_1, \ldots, \phi_D\}$, where each ϕ_i belong to a general space $\mathcal{H}$, and then represent each curve by means of a linear combination in $\mathrm{Span}(B)$. For this thesis we propose to consider $\mathcal{H}$ as a Reproducing Kernel Hilbert Space (RKHS) of functions see, Berlinet and Thomas-Agnan (2011); Muñoz and González (2010). In this case, the elements in the spanning set B are the eigenfunctions associated to the positive-definite and symmetric kernel function $K : T \times T \to \mathbb{R}$ that span $\mathcal{H}$. Therefore a functional data estimator of the discrete observation $\{x(t_i)\}_{i=1}^m$, denoted onwards as $\tilde{x}(t)$, is obtained by solving the following Support Vector Machine (SVM) regularization problem

$$\tilde{x}(t) := \arg\min_{f\in\mathcal{H}} \sum_{i=1}^{m} L_\varepsilon(x(t_i), f(t_i)) + \gamma\|f\|_{\mathcal{H}}^2, \tag{2.3}$$

where $\gamma > 0$ is a regularization parameter frequently chosen by cross–validation, $\|f\|_{\mathcal{H}}$ is the norm of the function f in $\mathcal{H}$ and $L(w, z) = (w - z)^2$ is a loss function. By the Representer Theorem Cucker and Smale (2002) the solution of the problem stated in Equation (2.3) exists, is unique, and admits the following representation

$$\tilde{x}(t) = \sum_{i=1}^{m} \alpha_i K(t, t_i) = \boldsymbol{\alpha}^T \mathbf{k}_t, \tag{2.4}$$

where $\mathbf{k}_t = (K(t_1, t), \ldots, K(t_m, t))$ is the vector of kernel evaluations and the linear combination coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m$ are obtained as the solution of the following linear system

$$(\gamma \mathbf{I}_m + \mathbf{K})\boldsymbol{\alpha} = \mathbf{y}, \tag{2.5}$$

for $\mathbf{y} = (x(t_1), \ldots, x(t_m))^T$, $\mathbf{I}_m$ an $m \times m$ identity matrix, and $\mathbf{K}$ the Gram matrix with the kernel function evaluations, $[\mathbf{K}]_{k,l} = K(t_k, t_l)$, for $k, l = 1, \ldots, m$. By the Mercer decomposition Theorem we have $K(t_k, t_l) = \sum_{j=1}^{\infty} \lambda_j \phi_j(t_k) \phi_j(t_l)$, where $(\lambda_j, \phi_j)_{j \geq 1}$ is a sequence of eigenvalue–eigenfunction pairs of the integral operator $I_K = \int_T K(\cdot, t) x(t) dt$, whose convergence is absolute and uniform by the Mercer Theorem, see J Mercer (1909). Combining this decomposition with the reproducing property, we can express each estimated functional datum $\tilde{x}(t)$ in the sample as follow

$$\tilde{x}(t) = \sum_{i=1}^{m} \sum_{j=1}^{\infty} \alpha_i \lambda_l \phi_l(t) \phi_l(t_i). \tag{2.6}$$

Nevertheless, the expression in Equation (2.6) is an unhelpful representation when the sequence of eigenpairs $(\lambda_j, \phi_j)_{j \geq 1}$ is unknown. Alternatively, the represented functional datum $\tilde{x}(t)$ can also be written using the eigenvalues $(l_1, \ldots, l_D)$ –in descending order– and the respective eigenvectors $(\mathbf{v}_1, \ldots, \mathbf{v}_D)$ of the rank-D Gram kernel matrix containing the kernel function evaluations –at points $t_1, \ldots, t_m$ and t–, and the representation of the functional datum is

$$\tilde{x}_D = \sum_{j=1}^{D} \sum_{i=1}^{m} \alpha_i l_j v_{i,j} v_{m+1,j} = \sum_{j=1}^{D} z_j e_j, \tag{2.7}$$

where $e_j = \sqrt{l_j} v_{m+1,j}$, $z_j = \dfrac{\sqrt{l_j}}{\sqrt{m}} \displaystyle\sum_{i=1}^{m} \alpha_i v_{i,j}$, and $D \leq m + 1$. In order to obtain a stable and low dimensional representation, Muñoz and González (2010), criteria such as the ratio $l_j / \sum_{j=1}^{D} l_j$, or the scree plot $\{(j, l_j)\}_{j=1}^{D}$, can be used in the practice to obtain a

suitable d-truncated representation as follows

$$\tilde{x}_d(t) = \sum_{j=1}^{d} \sum_{i=1}^{m} \alpha_i l_j v_{i,j} v_{m+1,j} = \sum_{j=1}^{d} z_j e_j, \text{ for } d < D \text{ and } \tilde{x}_d(t) \in \mathcal{H}_d \subset \mathcal{H}, \tag{2.8}$$

such that for a small ε_d it holds that $\sup_{t \in T} |\tilde{x}(t) - \tilde{x}_d(t)| \leq \varepsilon_d$. An **efficient representation for functional data identify each functional datum in the sample $\tilde{x}_s(t)$ with a vector $\mathbf{z}_s = (z_{1,s}, \ldots, z_{d,s}) \in \mathbb{R}^d$** for $s = 1, \ldots, n$.

o

Chapter 3

On the concept of order in a functional framework

In the era of Big Data, the statistical community pay special atention to the development of measures that induce an order in a complex data set, and in particular in a functional data context. Having an order in a data set is particulary useful to solve classification and outlier detection problems, among others.

In this sense order can be induced by what is called order statistics such as ranks and L-stistics. Following the definition of the Enciclopedia of Statistical Sicences an order statistics indicate the position of a certain value in a random variable. Let consider a sample of random variables $x_1, \ldots, x_n$. If the elements of the sample are ordered in terms of its magnitude, then the it can be expressed as $x_{1:n} \leq x_{2:n} \leq \cdots \leq x_{n:n}$ and called *order statistics*, see David and Nagaraja (2004).

Inside the family of order statistics are included the L-statistics, which are linear combinations of the former ones, namely of the form:

$$L_n = \sum_{i=1}^{n} c_{in} x_{i:n}, \tag{3.1}$$

where $c_{1n}, \ldots, c_{nn}$ are given constants (weights). Examples of L-statistics are the sample mean, $\bar{x} = n^{-1} \sum_{i=1}^{n} x_{i:n}$,, the range, $x_{n:n} - x_{1:n}$ and the median, $x_{Me} = x_{(n+1)/2:n}$ when n is odd and $x_{Me} = \frac{1}{2}(x_{n/2:n} + x_{n/2+1:n})$ when n is even. In particular when the x_i are independent and identically distributed it is considered a location statistic. For further examples see Boos (2004).

Besides the concept or order statistics a natural tool to analyze the order of the elements in a given data set is the concept of statistical depth function. In the rest of this chapter we deeply study this concept as a way to induce a total order in a given data set. First we introduce the concepts in a multivariate framework and then we study several extensions to the functional context and in particular we develop our own proposal of order for functional data.

3.1 Statistical depth function

Depth measures are often used in Statistics to order the data in a sample with respect to a location statistic, typically the mean, the median or the mode. Consider for instance a sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ where $\mathbf{x}_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$ are realizations of the random vector $\mathbf{X}$. When $d = 1$, to induce and order in the data we only need to consider the order statistics, say $x_{1:n}, x_{2:n}, \ldots, x_{n:n}$; then define as the central or deepest point the estimated median[1], namely $\hat{x}_{Me}$, and define the depth measure of a point x in the support of the distribution as $D(x, \hat{x}_{Me}) = |x - \hat{x}_{Me}|$.

In the multivariate context, when $d \geq 2$, the notion of center and order are not so clear as in the univariate case. A strategy commonly used when $d \geq 2$ consist in estimating a central or deepest point, for instance the sample mean, and then order the data according to its degree of centrality by ranking the observations with respect to some predefined metric from each point in the sample with respect to the estimated center. A well known example of this procedure is the Mahalanobis distance, commonly used when the sample data (approximately) follows a normal distribution. A function that implements that mapping is called in the literature, a 'depth function'. Therefore, a depth measure determine the degree of centrality –or outlyingness– of a point in a multivariate data set given an underlying distribution of the data at hand, see Liu et al. (1999); Zuo and Serfling (2000).

Up to now we have introduced the concept of depth. Formally in Zuo and Serfling (2000) the authors define a statistical depth function as follows:

Definition 3.1. Statistical depth function. Consider $\mathcal{F}$ the class of distributions on the Borel sets of $\mathbb{R}^d$ and $F_{\mathbf{x}}$ the distribution of $\mathbf{x} \in \mathbb{R}^d$. Then the bounded and non-negative mapping $D(\cdot;\cdot) : \mathbb{R}^d \times \mathcal{F} \to \mathbb{R}$ is a statistical depth function if satisfies the following properties:

[1] Is straightforward to prove that when $x \in \mathbb{R}$, the median is the most central or deepest point.

P1. **Affine invariance**. The depth of a given point does not change if an affine transformation is applied. $D(\mathbf{x}, F_{\mathbf{x}}) = D(A\mathbf{x} + b; F_{A\mathbf{x}+b})$.

P2. **Maximality at center**. The depth function should attain the maximum value at the center of the distribution (ussualy uniquely defined). $D(\mathbf{x}_0, F_{\mathbf{x}}) = \sup_{\mathbf{x}} D(\mathbf{x}; F_{\mathbf{x}})$, for any $\mathbf{x}_0 \in \mathbb{R}^d$ that is the center of $F_{\mathbf{x}}$.

P3. **Monotonicity relative to the deepest point**. As any point $\mathbf{x} \in \mathbb{R}^d$ turns away from the deepest point, the depth of $\mathbf{x}$ should decrease monotonically. $D(\mathbf{x}; F_{\mathbf{x}}) \leq D(\alpha\mathbf{x} + (1-\alpha)\mathbf{x}_0, F_{\mathbf{x}})$, for any $\mathbf{x}_0 \in \mathbb{R}^d$ that is the center of $F_{\mathbf{x}}$ and $\alpha \in [0, 1]$.

P4. **Vanishing at infinity**. For each $F_{\mathbf{x}},\ D(\mathbf{x}, F_{\mathbf{x}}) \to 0$, as $\|\mathbf{x}\| \to \infty$

Moreover Serfling (2006) mention four additional properties that are desirable but not necessary, and are listed below.

i *Symmetry*. Let $\mathbf{x}_0$ be the deepest point, if $F_{\mathbf{x}}$ is symmetric around $\mathbf{x}_0$, then so it is $D(\mathbf{x}; F_{\mathbf{x}})$.

ii *Continuity of* $D(\mathbf{x}, F_{\mathbf{x}})$ *as a function of* $\mathbf{x}$, (upper semicontinuity).

iii *Continuity of* $D(\mathbf{x}, F_{\mathbf{x}})$ *as a function of* $F_{\mathbf{x}}$.

iv *Quasi-concavity as a function of* $\mathbf{x}$. The set $\{\mathbf{x} : D(\mathbf{x}, F_{\mathbf{x}}) \geq c\}$ is convex for each real c.

3.2 Review of depth measures

In the following subsection the most widely used multivariate depth measures are presented, namely: i) the Mahalanobis depth, (MhD), ii) the half-space depth, (HD), iii) the simplicial depth (SD), iv) the Oja depth, (OD), and v) the Spatial depth, (SPD). For an extensive review of different depth measures and its properties, see Liu et al. (1999) an Zuo and Serfling (2000).

3.2.1 Multivariate depth measures

Let F be an absolutely continuous probability distribution in $\mathbb{R}^d$, with $d \geq 1$, and $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ a random sample of F, where each $\mathbf{x}_i$ is a column vector $d \times 1$. All the measures give the depth of a given point $\mathbf{x}$ relative to the distribution F.

Definition 3.2. Mahalanobis depth [Mahalanobis (1936)].

$$M_hD(\mathbf{x}, F) = [1 + (\mathbf{x} - \mu_F)\Sigma_F^{-1}(\mathbf{x} - \mu_F)^T]^{-1},$$

where μ_F and Σ_F are the mean vector and the covariance matrix of the distribution F. To obtain the sample version, μ_F and Σ_F must be substituted by their sample estimators.

Definition 3.3. Half-space (*Tukey*) depth [Tukey (1975)].

$$HD(\mathbf{x}, F) = \inf_H \{P(H) : \mathbf{x} \in H\},$$

where H is a closed halfspace in $\mathbb{R}^d$ and $\mathbf{x} \in H$. For the sample version F must be replaced by the empirical distribution F_n. The *Tukey* depth w.r.t a data set considers the minimum number of sample points of a distribution that belongs to one side of a hyperspace (halfspace) through the point $\mathbf{x}$.

Definition 3.4. Simplicial depth [Liu et al. (1990)].

$$SD(\mathbf{x}, F) = P_F\{\mathbf{x} \in S[\mathbf{x}_1, \ldots, \mathbf{x}_{d+1}]\},$$

where $S[\mathbf{x}_1, \ldots, \mathbf{x}_{d+1}]$ is a closed simplex of $(d + 1)$ random observations of F. The idea behind this measure is to construct all the possible simplices –triangles if $\mathbf{X} \in \mathbb{R}^2$–, and the deepest point will be the one that belongs to more simplices. The estiamted simplicial depth is as follows,

$$\widehat{SD}(\mathbf{x}, F_n) = \binom{n}{d+1}^{-1} \sum 1 \leq i_i, \ldots, i_d \leq n \mathbf{I}_{(\mathbf{x} \in S[\mathbf{x}_{i1}, \ldots, \mathbf{x}_{id+1}])}.$$

Definition 3.5. Oja depth [Oja (1983)].

$$OD(\mathbf{x}; F) = [1 + \mathbb{E}_F\{volume(S[\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_d])\}]^{-1},$$

where $S[\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_{d+1}]$ is a closed simplex with vertices $\mathbf{x}$, and $(d + 1)$ random observations of F. This measure computes the sum of the volume of every closed simplex with vertex in $\mathbf{x}$ and the others in any point of F. For $d = 2$ the Oja depth for a point $\mathbf{x}$ with respect to F is the sum of the areas of all the triangles which have one vertex at $\mathbf{x}$. the sample Oja depth is obtained by replacing F by the empirical distribution F_n:

$$\widehat{OD}(\mathbf{x}; F_n) = \binom{n}{d}^{-1} [1 + \sum_{1 \leq i_i, \ldots, i_d \leq n} \{volume(S[\mathbf{x}, \mathbf{x}_i, \ldots, \mathbf{x}_i])\}]^{-1}.$$

Definition 3.6. Spatial depth [Serfling (2002)].

$$SPD(\mathbf{x}; F) = 1 - \|\mathbb{E}[Sgn(\mathbf{x} - \mathbf{X})]\|,$$

where $\|\cdot\|$ is de Euclidean norm in $\mathbb{R}^d$ and $Sgn : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ is a sign function defined by:

$$S(\mathbf{x}) = \begin{cases} \dfrac{\mathbf{x}}{\|\mathbf{x}\|}, & \mathbf{x} \neq \mathbf{0}, \\ \mathbf{0}, & \mathbf{x} = \mathbf{0} \end{cases}$$

It is relevant to analyse whether these depth measures satisfy the properties 1-4. The *half-space* or *Tukey* depth satisfy the four properties described above. Also the *Mahalanobis* depth satisfy that four properties but only when F is symmetric. With respect to the *Simplicial* depth function, it also satisfies properties from 1-4 but for the case when F is an angularly symmetric[2] distribution, in other cases properties 2 and 3 are not always satisfied, see Zuo and Serfling (2000).

The median as the deepest point in $\mathbb{R}^d$

In a univariate context the median is well defined and is the most well known location measure of the center of a distribution as it is the value that splits the distribution in two equal parts; formally, for any real random variable with suppor $\mathbb{S} \subset \mathbb{R}$,

$$x_{Me} = arg \min_{s \in \mathbb{S}} \sum_{i=1}^{n} |x_i - s|,$$

the median is the value of the support of the distribution that minimize the sum of all the Euclidean distances between each value x_i and the rest of the elements in $\mathbb{S}$. In this univariate scenario the median presents the highest breakdown point, which is $\frac{(n-1)}{(2n)}$ and converges asymptotically to 0.5.

In finite dimensional Euclidean spaces ($\mathbb{R}^d,\ d \geq 2$) the median can be considered as the point in the support of the distribution with highest depth, according to a particular depth function. In that sense the multivariate depth based medians such as the half–space (Tukey) depth or the spatial depth do not satisfy the 50% breakdown point, and achieve $1/3$ and $1/(d+2)$ respectively, see Chakraborty and Chaudhuri (2014a). Besides the breakdown point issue, considering the multivariate median as the deepest point present some drawbacks when we deal with non-Guassian distributions, i.e.: assymetric or bimodal distributions. This is illustrated in Example 3.1.

[2] $\mathbf{X}$ present an angularly symmetric distribution in θ if $\frac{(\mathbf{X}-\theta)}{\|\mathbf{X}-\theta\|}$ is centrally symmetric in θ, which means $\mathbf{X} - \theta \overset{d}{=} \theta - \mathbf{X}$, see Liu et al. (1990) for further details.

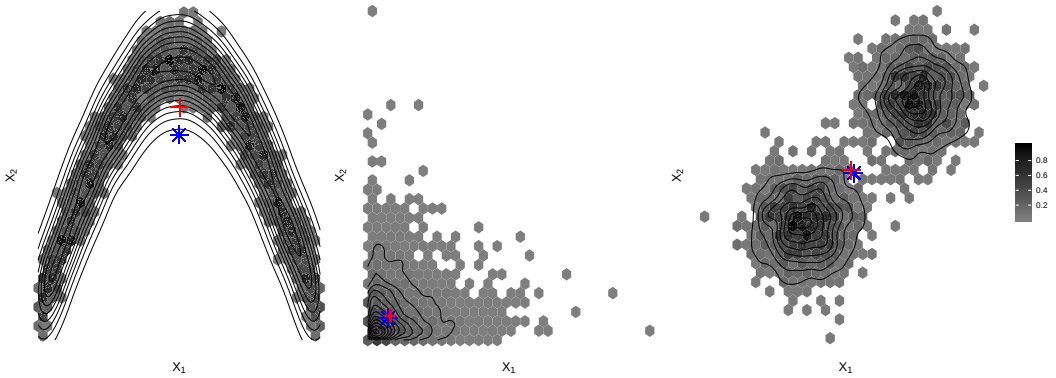


Figure 3.1: 2000 points in $\mathbb{R}^2$. Non-linear distribution (left), asymmetric scenario (center) and bi-modal scenario (rigth). In ("•") the coordinate-wise median and in red ("$*$") the Tukey, Half-space and spatial deepest point.

Example 3.1. The coordinate-wise median and median–based depth measures. In Figure 3.1 are illustrated three different scenarios. In the first one (left) the data was simulated from the following configuration: $x_2 = \sin(x_1) + \varepsilon$, where $x_1 \in [0, \pi]$ and $\varepsilon \sim \mathcal{N}(\mu_\varepsilon = 0, \sigma_\varepsilon = 0.05)$; generating the inverted U-shape. The second scenario (middle) was constructed considering two independen Chi–square distributions, with two degrees of freedom, $x_1, x_2 \sim \chi^2_2$. The last scenario (right) is a mixture of bivariate Gaussian distributions, $x_1, x_2 \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_{x_1} = (2, 2)$, $\boldsymbol{\mu}_{x_2} = (5, 5)$ and $\boldsymbol{\Sigma} = diag(0.05, 0.05)$

In each of the three scenarios is also depicted the coordinate–wise median in $\mathbb{R}^2$ (blue dot) and the deepest point obtained considering the Tukey, Half–Space and Spatial depth measures (red star). Two things are interesting to remark; i) in none of the three scenarios the coordinate–wise median belong to a high-density level set of the distributions. This show that under non-Gaussian scenarios the median is not a goode measure of centrality. Morover, in the first scenario the median does not respect the geometry embedded in the distribution. ii) The depth measures considered determine as deepest point an observation that does not belong to a high density region of the distribution. In this sence these multivariate location measures are not able to reflect information related to the center of the distribution. In general, under non-Gaussian scenarios any order induced considering a metric with respect to the median or using the depth measures mentioned, will be inadequate.

• • •

3.2.2 Notion of functional depth

In the case of infinite dimensional spaces as the case of the space of real–valued functions defined on a compact interval, namely functional data, the definition of a depth function and a functional median is much more complex. To begin let us define the concept of empirical functional median or coordinate–wise median and then reach to more sophisticated definitions.

Definition 3.7. Empirical functional median. Let $\{x_1(t_i), \ldots, x_n(t_i)\}_{i=1}^m,\ t \in T$, be a sample of random curves (raw functional data). The functional median is defined as:

$$x_{me}(t) \equiv \{(t_i, x_i^{me}) \in T \times \mathbb{R}\}_{i=1}^m, \tag{3.2}$$

where $x_i^{me} = \left\{ arg \min_{s \in \mathbb{R}} \sum_{l=1}^{n} |x_{il} - s| \right\}_{i=1}^{m}$.

The empirical functional median is constituted by the pair of points $\{t_i, x^{me}(t_i)\}$ such that for each t_i the correspondent value $x(t_i)$ is the univariate median. Considering this as the starting point, the concept of depth can be extended to infinite dimensional spaces in several ways. But, in any case, it has the same objective: measure the degree of centrality of a point, in this case a curve (discretzed function), with respect to a sample of functional data.

In the next subsections we describe the most widely used depth measures for functional data. To clarify the notation, for all the measures we consider a random sample of curves $\mathcal{S}_n = \{\tilde{x}_1(t), \ldots, \tilde{x}_n(t)\}$, $t \in T$ –as defined in Chapter 2–. We can assume that the $t_i's$ are common for all the curves. The depth measure for a curve $\tilde{x}(t) \in \mathcal{H}$ with respect to a set of curves $\mathcal{S}_n$ wil be denoted by $D(\tilde{x}(t), \mathcal{S}_n)$.

The band depth measure

The band-depth measure arose from a graph-based approach as a methodology to find the depth of an element in a given space with respect to a sample of functional data. It can be considered as a functional extension of the idea proposed in the simplicial depth by Liu et al. (1990). Consider a band in $\mathbb{R}^2$ delimited by the curves $\{\tilde{x}_l(t)\}_{l=1}^n$ as:

$$B(\tilde{x}_1(t), \ldots, \tilde{x}_l(t)) = \{(t, y) : t \in T, \min_{l=1,\ldots,n} \tilde{x}_l(t) \leq y \leq \max_{l=1,\ldots,n} \tilde{x}_l(t)\}.$$

The band depth measure for the function $\tilde{x}(t)$ is:

Definition 3.8. The band depth [López-Pintado and Romo (2009)].

$$BD(\tilde{x}(t), \mathcal{S}_n) = \sum_{l=2}^{n} Pr\bigg(G(\tilde{x}(t)) \subseteq B(\tilde{x}_1(t), \ldots, \tilde{x}_l(t))\bigg).$$

with $J \leq n$ and where $G(\tilde{x}(t)) = \{(t, \tilde{x}(t)) : t \in T\}$ is the graph of the funtion $\tilde{x}(t)$. The authors use $l = 2$ because i) $l > 3$ could be computationally expensive. The idea behind the band-depth measure is, given a set of curves compute all the possible bands defined by two curves. Then count all the curves that are included in each band. The curve that belongs to more bands is the deepest one. n that sense the band depth satisfy properties from 2 to 4, the affine invariance property is not satisfied. For a formal proof of these properties see, López-Pintado and Romo (2009).

The modified band depth measure

The modified band depth is a more flexible method to measure the depth of a curve given a functional data set. The indicator function is replaced by a measure of the "proportion" of the domain t that a curve is inside the band. That proportion is captured through the Lebesgue measure. Formally, for $2 \leq l \leq m$ and for any curve $\tilde{x}(t)$, let be $A_l(\tilde{x}(t))$, the interval in the domain T where $\tilde{x}(t)$ is inside the band formed by $B(\tilde{x}_1(t), \ldots, \tilde{x}_l(t))$,

$$A_l(\tilde{x}(t)) \equiv \bigg\{t \in T : \min_{r=1,\ldots,l} \tilde{x}_r(t) \leq \tilde{x}(t) \leq \max_{r=1,\ldots,l} \tilde{x}_r(t)\bigg\}.$$

Then a measure of the time that this occurs is proposed by the ratio $\lambda_r(A_l(\tilde{x}(t))) = \dfrac{\lambda(A_l(\tilde{x}(t)))}{\lambda(T)}$. Let $l \in [2, n]$. The modified band depth measure of $\tilde{x}(t)$ is:

Definition 3.9. The modified band depth [López-Pintado and Romo (2009)].

$$MBD(\tilde{x}(t), \mathcal{S}_n) = \sum_{j=2}^{J} \mathbb{E}\bigg[\lambda_r(A_l(\tilde{x}(t)))\bigg].$$

For this version the authors consider $l = 2$ because it is computationally fast and also the results are stable with respect to l. The idea behind the modified version of the band depth is for a sample of n curves, consider 'bands' defined for combinations of 2 curves, and account for the "proportion" of the domain T that a curve $\tilde{x}(t)$ is contained in the band (depth index). Hence, the depth of $\tilde{x}(t)$ is defined as the average of the depth index for all the possible bands. The deepest curve is the curve with the maximum depth.

The random Tukey depth

The random Tukey depth (RTD), is a random approximation of the Tukey depth or halfspace depth. Consider a separable metric space $(\mathcal{F}, d) = (\mathcal{H}, \|\cdot\|_{L_2})$, where $\mathcal{H}$ is an infinte-dimensional Hilbert space, and define $\mathcal{U} = \{u_1, \ldots, u_k\}$ each one sampled independently from a nondegenerate probability measure μ in $\mathbb{H}$. $\mathcal{S}_n$ is a set of functions (curves) defined as a functional random variable on the probability space $(\mathcal{F}, \mathcal{A}, P)$, where $\mathcal{A}$ is the Borel sigma algebra and P is a probability measure on the Borel sets of $\mathcal{A}$. The random Tukey depth for a function $\tilde{x}(t)$ with respect to a set of curves $\mathcal{S}_n$ is:

Definition 3.10. The random Tukey depth [Cuesta-Albertos and Nieto-Reyes (2008)].

$$RTD(\tilde{x}(t), \mathcal{S}_n) = \min_{u \in \mathcal{U}} D_1(\langle u, c\rangle, P_u),$$

where P_u is the marginal of the probability distribution P and for each probability measure Q in a Borel set $\mathbb{R}$, $D_1(t, Q) = min\{Q(-\infty, t], Q[t, -\infty)\}$. The sample version is obtained by substituting P by P_n. This depth function is a random variable in itself, then for the same functional data set can take different values, and then order the data in different ways. For further details see Cuesta-Albertos and Nieto-Reyes (2008).

The h-mode depth

The h-mode depth considers the average of the kernelized distances using the L_2 norm. The h-mode depth for a function $\tilde{x}(t)$ with respect to S_n is:

Definition 3.11. The h-mode depth [Cuevas et al. (2007)].

$$h - MD(\tilde{x}(t), \mathcal{S}_n) = \frac{1}{nh} \sum_{l=1}^{n} K\left(\frac{\|\tilde{x}(t) - \tilde{x}_l(t)\|_{L_2}}{h}\right) = \sum_{l=1}^{n} K_h(\|\tilde{x}(t) - \tilde{x}_l(t)\|),$$

where K_h is a kernel function such that $K_k = \frac{1}{h} K(\tilde{x}(t), \mathcal{S}_n)$ and h is a fixed tunning parameter (bandwidth). For further details and consistency proofs see Cuevas et al. (2007) and Nagy (2015).

The functional spatial depth

The functional spatial depth can be expressed in the same way as the spatial depth (see Def. 3.6), adapting the norm and the sign function to one suitable for a separable hilbert space.

Definition 3.12. The functional spatial depth [Chakraborty and Chaudhuri (2014b)].

$$FSD(\tilde{x}(t), \mathcal{S}_n) = 1 - \|\mathbb{E}[FS((\tilde{x}(t) - \mathcal{S}_n)]\|,$$

where $\|\cdot\|_{L_2}$ is the L_2 norm and $FS : \mathcal{H} \longrightarrow \mathcal{H}$ is a functional sign function defined by:

$$FS(\tilde{x}(t)) = \begin{cases} \dfrac{\tilde{x}(t)}{\|\tilde{x}(t)\|_{L_2}}, & \tilde{x}(t) \neq 0, \\ 0, & \tilde{x}(t) = 0 \end{cases}$$

Other functional depths

In the literature there are other several contributions to the concept of functional depth. For instance, the Fraiman-Muniz (Integrated) depth proposed by Fraiman and Muniz (2001) measures the conditional quantile on all points. Moreover when the modified band depth is computed with $l = 2$, which is the value used in López-Pintado and Romo (2009), this measure and the integrated depth coincide Nieto-Reyes and Battey (2016). The half-region and modified half-region depth, López-Pintado and Romo (2011), are constructed taking into account the hypograph and epigraph of a (curve). It computes the proportion of curves whose graph belongs to the hypograph of $\tilde{x}(t)$, and the epigraph of $\tilde{x}(t)$, and then take the minimum value. The kernel functional spatial depth (KFSD) Sguera et al. (2014) is the kernelized version of the functional spatial depth.

Some drawbacks

The aforementioned depth measures suffer from drawbacks: i) some of these methodologies when dealing with depth meausres for functional data, use the raw representation of the data, for instance $(t_i, x(t_i))$ in the case of a set of univaraite time series, ignoring the functional nature of the data, (see López-Pintado and Romo (2009, 2011); Fraiman and Muniz (2001)). ii) The computation of these metrics requires evaluating a large number of integrals, which is impractical when working with many curves. iii) Some of this measures, such as the random Tukey depth –that reduce computational cost by evaluating projections– do not provide a stable criterion to order the functions. v) To determine which function is atypical it is essential to establish a probabilistic threshold to make such decision, which is not possible in this measures. v) In the case of the measures BD, MBD, Integrated, HRD y MHRD, the empirical functional median is the curve with highest depth, see Chakraborty and Chaudhuri (2014a) and Appendix A for a formal proof. This can can lead to inaccurate results when there are non linearities in the functional space where the curves are defined. In other words, a functional depth should respect the geometry of the manifold embedded in $\mathcal{H}$. An illustration of this is

presented in Example 3.2.

Some of these drawbacks are tackled in this chapter through the proposal of a new functional depth measure: the *Generalized Kernel Depth* (GKD), detailed in the next section.

3.3 Kernel depth measures for functional data

A natural way to order a data set is to consider a centrality measure or deepest point and a metric– for instance the Euclidean distance–, so to compute the distance between each point in the sample with respect to the center, and then sort the data by means of this metric. Thus, depth functions compute how deep is a point with respect to a distribution/data set, defining its degree of 'centrality' or 'outlyingness'. Hence, a proper estimation of the deepest point must be done.

In line with the above, we think that a functional measure of centrality or depth must:

i) reflect information related to the center of the underlying distribution of the generating process of the functional data (curves);

ii) respect the geometry of the distribution of that generating process.

To tackle point i) we start defining the center of a distribution.

Definition 3.13 (Center of a distribution)**.** Let Z be a random vector with distribution F, we say $\mathbf{m} \in \text{support}(Z)$ is the center of the distribution if $P(Z \in B(\mathbf{m}, \varepsilon)) = \int \mathbb{I}_{||Z-\mathbf{m}||\leq\varepsilon}(Z)dF \geq P(Z \in B(\mathbf{z}, \varepsilon))$ for all $\mathbf{z} \in \text{support}(Z)$, where $B(\mathbf{z}, \varepsilon)$ is a ball with center in $\mathbf{z}$ and –sufficiently small– radius ε.

Several location measures satisfy the requirements of Definition 3.13 to be the center of a distribution; for example the mean in the family of Gaussian distributions. In practice, we do not know the distribution of the random elements at hand, therefore an estimator for the center is needed.

Definition 3.14 (Estimated kernel center)**.** Given a random sample of curves $\mathcal{S}_n = \{\tilde{x}_1(t), \ldots, \tilde{x}_n(t)\}$ –realizations of the stochastic process $X(\omega, t)$–, and its corresponding truncated $\mathcal{H}_d$-representations, namely $\mathbf{z}_s = (z_{1,s}, \ldots, z_{d,s}) \in \mathbb{R}^d$ for $s = 1, \ldots, n$; we denote by $\hat{P}_n$ to any consistent estimator of the distribution of the random vector Z, and define the kernel center as $\widehat{\mathbf{m}} \in \text{support}(Z)$ such that $\hat{P}_n(Z \in B(\widehat{\mathbf{m}}, \varepsilon)) \geq \hat{P}_n(Z \in B(\mathbf{z}, \varepsilon))$, for all $\mathbf{z} \in \text{support}(Z)$ and sufficiently small ε.

Diverse center estimators can be used in practice. For instance, we could estimate the deepest functional data point $\widehat{\mathbf{m}}$ as the multivariate median of the d-dimensional data points $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$, computed as the vector of the coordinatewise medians in $\mathbb{R}^d$, or what can be called the vector of marginal medians. That is $\widehat{\mathbf{m}} = (m_1, \ldots, m_d)$, where $m_i = \text{median}\{z_{i,1}, \ldots, z_{i,n}\}$ for $i = 1, \ldots, d$. Following the same road as in Liu et al. (1999); Hu et al. (2011), next we define the Kernel Mahalanobis Depth.

Definition 3.15 (Estimated Kernel Mahalanobis Depth)**.** Given a random sample of curves $\mathcal{S}_n = \{\tilde{x}_1(t), \ldots, \tilde{x}_n(t)\}$, the estimated Kernel Mahalanobis Depth (KMD) of a functional datum $\tilde{x}(t) \in \mathcal{H}_d$ is defined as

$$KMD(\tilde{x}(t), \mathcal{S}_n) := [(\mathbf{z} - \widehat{\mathbf{m}})^T \widehat{\mathbf{\Sigma}}^{-1} (\mathbf{z} - \widehat{\mathbf{m}})]^{-1/2}, \tag{3.3}$$

where $\widehat{\mathbf{m}}$ is the estimated kernel centrality measure and $\widehat{\mathbf{\Sigma}}$ is the inverse of the sample covariance matrix, both estimated using $\{\mathbf{z}_1 \ldots, \mathbf{z}_n\}$.

3.3.1 The Generalized kernel depth

The use of depth measures based on the Mahalanobis distance, as in Equation (3.3), present two main drawbacks Hu et al. (2011); Martos et al. (2014); Zhang et al. (2012): i) relay on the strong assumption that the underlying distribution of the data is Gaussian and ii) to compute the distance an estimation of the first two moments is needed, which can lead to problems when the intrinsic dimensionality of the data increases or there are outliers in the sample. To solve these problems, and following the methodology defined by the authors in Martos et al. (2014), we propose a generalization of the Mahalanobis depth via density kernels. This generalization involves defining a family of kernels based on the underlying density function of the data at hand. Previously is important to introduce the concept of asymptotic f–monotonicity.

Definition 3.16 (Asymptotic f-monotonicity)**.** Consider a random sample $Z_n = \{\mathbf{z}_i\}_{i=1}^n$ drawn from a probability distribution F and denote by f_F $Z \mapsto \mathbb{R}^+$ the corresponding density function. A function $g(\mathbf{z}, Z_n)$ is asymptotically f-monotone if:

$$f_F(\mathbf{z}) \geq f_F(\mathbf{y}) \Rightarrow \lim_{n \to \infty} P(g(\mathbf{z}, Z_n) \geq g(\mathbf{y}, Z_n)) = 1. \tag{3.4}$$

Definition 3.17 (Density Kernel)**.** Let Z be a random vector in $\mathbb{R}^d$ distributed according to a measure F that admits a bounded probability density function f and let $g(\mathbf{z}, F)$ be a positive f-monotone function. Define $\phi_F : Z \to \mathbb{R}^+$ as $\phi_F(\mathbf{z}) = g(\mathbf{z}, F)$. The density kernel is defined as:

$$K_F(\mathbf{z}, \mathbf{y}) = \phi_F(\mathbf{z}) \phi_F(\mathbf{y}) \tag{3.5}$$

Next we propose the generalized kernel depth measure.

Definition 3.18 (Generalized Kernel Depth)**.** Given a random sample of curves $\mathcal{S}_n = \{\tilde{x}_1(t), \ldots, \tilde{x}_n(t)\}$ and a density Kernel K_F, the Generalized Kernel Depth (GKD) of curve $\tilde{x}(t) \in \mathcal{H}_d$, represented with the coefficients $\mathbf{z}$, is defined as follows:

$$GKD(\tilde{x}(t), \mathcal{S}_n, K_F) = \phi_F(\mathbf{z})\phi_F(\widehat{\mathbf{m}}_0), \tag{3.6}$$

where $\mathbf{m}_0 = \arg\max_{\mathbf{z}} f(\mathbf{z})$ is the center of the distribution F –one or more points in the support of Z–, and $\widehat{\mathbf{m}}_0$ is the estimated mode using $\{\mathbf{z}_1 \ldots, \mathbf{z}_n\}$. The GKD has the desirable properties of a well defined depth measure (see Zuo and Serfling (2000)), which are stated in Proposition 3.1.

Proposition 3.1. *For the GKD as defined in 3.18, the following properties are satisfied.*

P1. **Maximality at center***: Let $\widetilde{m}_0(t)$ be the estimated modal curve –that is the curve in $\mathcal{H}_d$ identified with $\widehat{\mathbf{m}}_0$–, then it holds that*

$$GKD(\widetilde{m}_0(t), \mathcal{S}_n, K_F) = \sup_{z \in \mathbb{R}^d} \phi_F(\mathbf{z})\phi_F(\widehat{\mathbf{m}}_0)$$

P2. **Monotonicity relative to the deepest point***: For any $\tilde{x}(t) \in \mathcal{H}_d$*

$$GKD(\tilde{x}(t), \mathcal{S}_n, K_F) \leq GKD(\widetilde{m}_0(t), \mathcal{S}_n, K_F).$$

P3. **Vanishes at infinity***: For $\mathbf{z}$ representing $\tilde{x}(t) \in \mathcal{H}_d$, it holds that*

$$GKD(\tilde{x}(t), \mathcal{S}_n, K_F) \to 0 \text{ if } ||\mathbf{z}|| \to \infty.$$

P4. **Invariant under affine transformations***: Let $\mathcal{T}$ be the class of affine transformations in $\mathcal{H}_d$ and let $\tau \in \mathcal{T}$ be an affine map, then*

$$GKD(\tilde{x}(t), \mathcal{S}_n, K_F) = GKD(\tau \circ \tilde{x}(t), \mathcal{S}_n, K_F).$$

P5. **Invariant to RKHS representation***: The order induced on the sample of functional data is not altered when a different functional basis –i.e. kernel parameters– is chosen.*

To obtain an asymptotically optimal representation, we choose the kernel and regularization parameters of Equation (2.3) by cross–validation. In this way, Property P.5, is of fundamental importance since ensures that the center-outward ordering induced by GKD is independent of these parameters choice (see Appendix A for a formal proof).

Proposition 3.2. *The Mahalanobis depth as defined in Definition 3.2 is a particular case of the GKD.*

Proposition 3.3. *The h-mode depth as defined in Definition 3.11 is a particular case of the GKD.*

For a formal proof of Propositions 3.2 and 3.3 see Appendix A.

3.3.2 Estimating the GKD

The estimation procedure of the GKD for a curve x_s with respect to a sample of raw–functional data X is detailed in Algorithm 1.

Algorithm 1: Estimation of $GKD(x, X, K)$ from a sample of raw functional data.

GKD: $(x_s, X, K, \gamma, d, density)$;
Input : The curve x_s, the raw–functional data matrix $X \in \mathbb{R}^{n\times m}$ –paths in rows–, the kernel function K, the regularization parameter γ, the truncation parameter $d \leq \text{rank}(\mathbf{K})$, and a predefined *density* estimation procedure.
Output: $GKD(x, X, K)$
for *l in 1 to n* **do**
  compute $\boldsymbol{a}_l = (\gamma m\mathbf{I} + \mathbf{K})^{-1}\mathbf{x}_l$;
  for *j in 1 to d* **do**
    $\widehat{\xi}_{l,j} = \sqrt{\lambda_j}\sum_{i=1}^{m} a_{l,i}v_{i,j}$
  end
  store $\mathbf{z}_l = (\widehat{\xi}_{1,l}, \ldots, \widehat{\xi}_{d,l})$
end
Consider $S_n = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ an iid sample from the random vector Z. Estimate $\widehat{F}_Z$ with a predefined *density estimation* procedure and compute $\mathbf{m}_0 = \arg\max_{\mathbf{z}} f(\mathbf{z})$. Return $GKD(x, X, K) = \phi_F(\mathbf{z}_s)\phi_F(\mathbf{m}_0)$.

We propose de use of the following expression for the function $\phi_F(\mathbf{z}) = \dfrac{f(\mathbf{z})}{f(\mathbf{m}_0)}$. Following this configuration the Estimated GKD is presented in the next definition:

Definition 3.19 (Estimated Generalized Kernel Depth)**.** Given a random sample of curves $\mathcal{S}_n = \{\tilde{x}_1(t), \ldots, \tilde{x}_n(t)\}$ and a density Kernel K_F, the estimated Generalized Kernel Depth ($\widehat{GKD}$) of curve $\tilde{x}(t) \in \mathcal{H}_d$, represented with the coefficients $\mathbf{z}$, is defined as follows:

$$\widehat{GKD}(\tilde{x}(t), \mathcal{S}_n, K_F) = \frac{f(\mathbf{z})}{f(\mathbf{m}_0)}, \tag{3.7}$$

A wide family of density estimators can be aaplied to compute the Estimated GKD. As this chapter focus the order induced by the depth function and to avoid the computational costs of estimating the density, we suggest to estimate the α–volume sets of $f(\mathbf{Z})$, which are defined by $V_\alpha(f) = \{z \in Z | f(z) \geq \alpha\}$, such that $P(V_\alpha(f)) = 1 - \nu$, where $0 < \nu < 1$. To estimate the V_α we consider the Once–Class Neighbor Machine (OCNM), see Moguerza and Muñoz (2006).

Example 3.2. The empirical functional median as the deepest curve. Three different scenarios of simulated functional data are presented in Figure 3.2. The simulations were obtained by the following generating processes

$$\tilde{x}(t) = \xi_1 \sin(\pi t) + \xi_2 \sin(2\pi t)$$

where $t \in [0, 1]$. For the first scenario (upper panel) (ξ_1, ξ_2) are distributed according to the following configuration: $\xi_2 = \sin(\xi_1) + \varepsilon$, where $\xi_1 \in [0, \pi]$ and $\varepsilon \sim \mathcal{N}(\mu_\varepsilon = 0, \sigma_\varepsilon = 0.05)$ –generating an inverted U-shape–. The second scenario (middle panel) was constructed considering two independent Chi–square distributions, with two degrees of freedom, $\xi_1, \xi_2 \sim \chi^2_2$. The last scenario (bottom panel) is a mixture of bivariate Gaussian distributions, $\xi_1, \xi_2 \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_{\xi_1} = (2, 2)$, $\boldsymbol{\mu}_{\xi_2} = (5, 5)$ and $\boldsymbol{\Sigma} = diag(0.05, 0.05)$.

As can be appreciated in Figure 3.2, when we analyze the original representation of the functional data, that is in the coordinates $(t_i, x(t_i))$, the empirical functional median, computed as in Definition 3.7, is located in the center of the data. Analyzing the distribution of the coefficients of the underlying genrating process of the functions, we can realize that what seems to be reasonable really is not. The functional median does not represent a function correspoinding with a point –in the space spanned by the first two functional principal components– of high density. Besides, as it is cleary shown in the first scenario –upper right chart–, the functional median does not respect the geometry of the distrubition of the realizations of the generating process.

The GKD outperforms the empirical functional median in task of defininig the deepest point of a functional data set. In that sense the deepest function identified by the GKD –in the three scenarios– satisfy the two characteristics stated at the begining of this Section: i) reflects information related to the center of the underlying distribution of the generating process of the functional data and ii) respect its geometry. Even in the case of the bi–modal scenario wher the GKD present two local maxima.

• • •

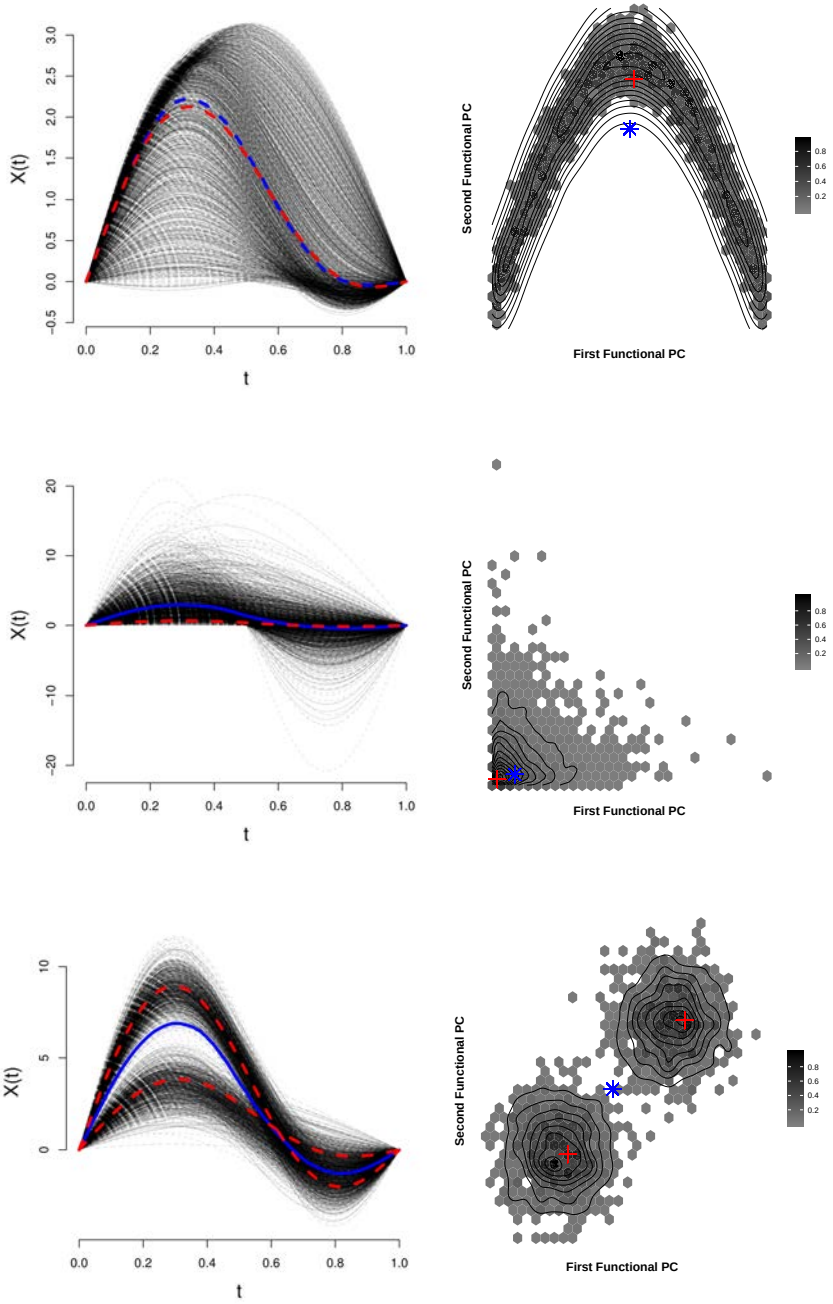


Figure 3.2: 2000 curves (left) and the corresponding 1^{st} and 2^{nd} functional principal component (right). In the upper panel the Non-linear configuration. The asymmetric scenario (center) and the bi-modal scenario (bottom). In ("$---$") the empirifcal functional median and in ("$---$") the GKD deepest curve(s). In blue ("$*$") and ("$+$") its corresponding first and second functional principal component respectively.

3.3.3 Using KMD and GKD measures for functional outlier identification

Because depth induces a center-outward ordering of a sample of functional data, a natural problem to test the utility of different depth definitions is outlier detection. To obtain a usefull definition of the GKD, we just need and estimation of the density function $f_{\mathbb{F}}$ of the random sample $Z_n = \{\mathbf{z}_i\}_{i=1}^n$ and then we obtain the relative order of each point with respect to the mode of the distribution $\mathbf{m}_0$.

The KMD measure in Definition 3.15 is computed as the inverse of the Mahalanobis distance in $\mathcal{H}_d$; therefore seems natural to consider the Multivariate Normal distribution as the most suitable probability model to determine whether a curve in the sample of raw functional data is an outlier. To this aim, we proceed as follows: estimate $(\widehat{\mathbf{m}}, \widehat{\mathbf{\Sigma}})$ –using the n curves in the sample represented with $\mathbf{z}_1, \ldots, \mathbf{z}_n$–, and then define

$$\tilde{x}(t) \text{ is an outlier if } \text{KMD}(\tilde{x}(t), \mathcal{S}_n) < \left(\chi^2_d(\nu)\right)^{-1/2},$$

where $\chi^2_d(\nu)$ is the $1-\nu$ quantile of a Chi-square distribution with d-degrees of freedom. When the proportion of outlier in the sample, denoted as ν onwards, is known a priori, the $\chi^2_d(\nu)$-quantile threshold can be replaced by the corresponding $(1-\nu)$ sample quantile of the KMD–distance; that is, we classify as anomalous data the νn curves that are most KMD–distant to the center. When ν is unknown, the ratio $\text{KMD}(\tilde{x}_s(t), \mathcal{S}_n) / \sum_{s=1}^n \text{KMD}(\tilde{x}_s(t), \mathcal{S}_n)$, or the scree plot $\{(s, \text{KMD}(\tilde{x}_s(t), \mathcal{S}_n)\}_{s=1}^n$ are alternative criteria that could be used in practice to guide the selection of outlier curves in the sample.

As we are performing an outlier identification analysis, it can be assumed that a proportion $\nu > 0$ exists in the sample data. In this context the parameters of the KMD $(\widehat{\mathbf{m}}, \widehat{\mathbf{\Sigma}})$, will be affected by the presence of atypical observations. Consequently the KMD estimation can be robutified by estimating the parameters $(\mathbf{m}, \mathbf{\Sigma})$ with a robust maximum likelihood method. In the experimental section we present both results: KMD and RML-KMD.

With respect to the GKD, there is no limiting distribution for this depth measure, therefore different approaches could be implemented when using this metric to determine outliers in the sample of functional data. As in the case of KMD, when ν is known a priori, we can sort the curves according to the GKD and determine consider the curves in the $(1-\nu)$ GKD–quantile as the outliers in the sample. When the number of abnormal curves in the sample is unknown, we recommend to conduct a sensitivity analysis

over the parameter ν to determine its precise value.

3.4 Experimental work

In this section we perform numerical experiments to address the performance of the proposed kernel depths on the task of functional outlier detection. In what follows, when representing functional data, we consider the Gaussian Kernel function $K(t_k, t_l) = e^{-\sigma\|t_k - t_l\|^2}$. The penalization coefficient γ of the SVM regularization problem, was obtained throught cross validation, Muñoz et al. (2018).

In the task of identifying abnormal realizations of an stochastic process, we test our depth measures against several well known depth functions, namely: the modified band depth (MBD) López-Pintado and Romo (2009) already implemented in the R-package `'depthTools'` Lopez-Pintado and Torrente (2013), the random Tukey depth (RTD) and the h-mode depth (h-MD), see Cuevas et al. (2007); Cuesta-Albertos and Nieto-Reyes (2008), implemented in the R-package `'fda-usc'` Febrero-Bande and Oviedo de la Fuente (2013), and the functional spatial depth (FSD), see Chakraborty and Chaudhuri (2014b). For the real–data context, in the univariate functional data case, we also consider as competitor procedures the ones defined in Hyndman (1996) and Arribas-Gil and Romo (2014).

In the experiments developed in sections 3.4.3 and 3.4.4, which can be circumscribed under a problem of outlier identification for multivariate functional data, we test our mesure against the multivariate extension of the Modified Band Depth measure (MMBD) Ieva and Paganoni (2013) already implemented in the R-package `'roahd'` Tarabelloni (2018), and the multivariate functional projection depth (MFPD) Zuo et al. (2003) implemented in the R-package `'mrfdepth'` Segaert (2018). The depth measures proposed in this chapter are already implemented in the R-package `'bigdatadist'` (see Martos and Hernández (2018)).

3.4.1 Univariate functional data for Monte Carlo study

In this experiment we consider a random sample of $n = 400$ paths $\{x_1(t), \ldots, x_n(t)\}$, where a small proportion $\nu \in [0, 1]$, known a priori, of these paths present an atypical pattern, and the remaining $n(1 - \nu)$ curves are considered the main data. Through a Monte Carlo study, we test the performance of the proposed methods over three data configurations (scenarios A, B and C) and for three different values of the parameter

$\nu \in \{1\%, 5\%, 10\%\}$. Specifically, we consider the following generating processes:

$$\begin{aligned} X_l(t) &= \sum_{j=1}^{4} \xi_j \sin(j\pi t) + \varepsilon_l(t), \text{ for } l = 1, \ldots, (1-\nu)n, \\ Y_l(t) &= \sum_{j=1}^{4} \zeta_j \sin(j\pi t) + \varepsilon_l(t), \text{ for } l = 1, \ldots, \nu n/2, \\ Z_l(t) &= \sum_{j=1}^{4} \eta_j \sin(j\pi t) + \varepsilon_l(t), \text{ for } l = 1, \ldots, \nu n/2, \end{aligned}$$

where $t \in [0, 1]$ and $\varepsilon(t)$ are independent autocorrelated random error functions, for scenarios

(A) **Gaussian scenario:** $(\xi_1, \ldots, \xi_4)$ is a normally-distributed multivariate random variable (NDMRV) with mean $\boldsymbol{\mu}_\xi = (4, 2, 4, 1)$ and diagonal co-variance matrix $\boldsymbol{\Sigma}_\xi = \mathrm{diag}(5, 2, 2, 1)$. To generate magnitude outliers, we consider $(\zeta_1, \zeta_2, \zeta_3, \zeta_4)$ NDMRV with parameters $\boldsymbol{\mu}_\zeta = 2.5\boldsymbol{\mu}_\xi$ and $\boldsymbol{\Sigma}_\zeta = (2.5)^2\boldsymbol{\Sigma}_\xi$. To generate shape outliers, we choose $(\eta_1, \eta_2, \eta_3, \eta_4)$ NDMRV with parameters $\boldsymbol{\mu}_\eta = (4, -2, 1, 3)$ and $\boldsymbol{\Sigma}_\eta = \boldsymbol{\Sigma}_\xi$.

(B) **Asymmetric scenario:** $(\xi_1, \ldots, \xi_4)$ are independent Chi-square distributed r.v. (ICRV) with $16, 16, 12, 12$ degrees of freedom respectively; $\zeta_1, \zeta_2, \zeta_3, \zeta_4$ are ICRV with $40, 40, 30, 30$ degrees of freedom respectively; and $\eta_1, \eta_2, \eta_3, \eta_4$ are NDMRV with $\boldsymbol{\mu}_\eta = (18, 16, 8, -10)$ and $\boldsymbol{\Sigma}_\eta = \mathrm{diag}(15, 12, 12, 15)$.

(C) **Bi-modal scenario:** in this scenario we state a bi-modal distribution for the vector parameter $\boldsymbol{\xi}$. Let $b \sim \mathcal{B}((1-\nu)n, p)$ be a binomial random variable with parameter $p = 0.5$, then $(\xi_1, \ldots, \xi_4)$ is a NDMRV with mean $\mu_\xi = b(1, 1, 1, 1) + (1-b)(9, 9, 9, 9)$. That is, when $b = 0$ then $\boldsymbol{\mu}_\xi = (1, 1, 1, 1)$; otherwise $\boldsymbol{\mu}_\xi = (9, 9, 9, 9)$. For $b = \{0, 1\}$, the diagonal co-variance matrix is $\boldsymbol{\Sigma}_\xi = \mathrm{diag}(5, 2, 2, 1)$.

To generate outliers, we consider i) $(\zeta_1, \zeta_2, \zeta_3, \zeta_4)$ NDMRV with parameters $\boldsymbol{\mu}_\zeta = 2\boldsymbol{\mu}_\xi$ and $\boldsymbol{\Sigma}_\zeta = (2)^2\boldsymbol{\Sigma}_\xi$; and ii) $(\eta_1, \eta_2, \eta_3, \eta_4)$ NDMRV with parameters $\boldsymbol{\mu}_\eta = (5, 5, 5, 5)$ and $\boldsymbol{\Sigma}_\eta = 0.5\,\mathrm{diag}(1, 1, 1, 1)$.

To illustrate the generating process, in Figure 3.3, we show one instance of the simulated paths in Scenarios A, B and C –left, center and right respectively– with $\nu = 10\%$. For this experiment, the values of the parameter ν are assumed known in each scenario. With respect to the kernel parameter σ, and the penalization parameter of the regularization problem in Equation 2.6, we select them with a 10-fold cross-validation

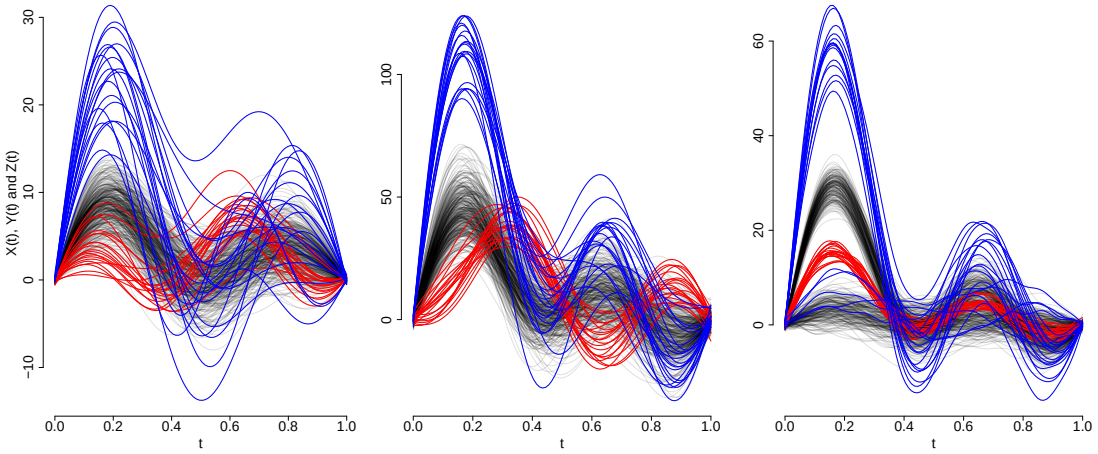


Figure 3.3: functional data, 400 curves corresponding to $\nu = 10\%$, Gaussian scenario (left), Asymmetric scenario (center) and Bi-modal scenario (rigth). In black ("—"), the sample of regular paths $X(t)$, and abnormal curves $Y(t)$ in red ("—"), and $Z(t)$ in blue ("—").

procedure using a single set of data, which correspond to the first instance of the simulations. The reference values, which remain fixed throughout the simulation exercise, are $\sigma = 10$ and $\gamma = 0.1^5$.

Let P and N be the amount of outlier and normal data in the sample, respectively, and let TP = True Positive and TN = True Negative be the respective quantities detected by different methods; in Table 3.1, we report the following average metrics TPR = TP/P (True Positive Rate or sensitivity), TNR = TN/N (True Negative Rate or specificity) and the area under the ROC curve (aROC) of each method obtained through the $M = 1000$ replications in the Monte Carlo study.

The KMD and GKD measures proposed in this work outperform the competitor depth measures in the three scenarios considered when $\nu \in \{1\%, 5\%, 10\%\}$. However in the case where $\nu = 1\%$, the standard errors are slightly high to confirm a significant difference between the methods.

When we compare among the proposed methods, the GKD seems to be slightly but consistently more effective than the KMD and robust–KMD in Scenario B and C where the distributions of the coefficients are asymmetric and multimodal respectively. In particular the difference is greater in scenario C given that the KMD is especially adequate for unimodal data, while the GKD method does not requires unimodal distributions to work. This result of empirical independence of the data distribution, is very interesting if we consider that in general the underlying distribution of the data is unknown.

Table 3.1: Simulation analysis: Scenarios and contamination percentages ν in columns. In rows, different methods and average sensitivities, specificities and the areas under the ROC curves (aROC) (this last on a scale of 10^2). The corresponding standard-error is reported in parenthesis.

Method	Metric	Scenario A			Scenario B			Scenario C		
		10%	5%	1%	10%	5%	1%	10%	5%	1%
MBD	TPR	67.963	58.575	36.125	81.203	74.380	52.200	30.545	27.585	24.425
		(5.270)	(7.972)	(18.431)	(4.269)	(7.076)	(21.286)	(5.206)	(7.413)	(17.374)
	TNR	96.440	97.820	99.355	97.911	98.651	99.517	92.283	96.189	99.237
		(0.586)	(0.420)	(0.186)	(0.474)	(0.372)	(0.215)	(0.578)	(0.390)	(0.175)
	aROC	95.310	95.409	95.506	98.443	98.498	98.543	58.496	58.840	57.963
		(1.376)	(1.847)	(3.849)	(0.576)	(0.732)	(1.444)	(3.167)	(4.380)	(10.411)
h-MD	TPR	80.078	76.770	67.000	83.048	79.840	70.475	79.088	79.240	67.175
		(4.414)	(6.860)	(18.305)	(4.272)	(6.145)	(16.098)	(5.752)	(6.994)	(18.663)
	TNR	97.786	98.777	99.667	98.116	98.939	99.702	97.676	98.907	**99.689**
		(0.586)	(0.420)	(0.186)	(0.474)	(0.372)	(0.215)	(0.578)	(0.390)	(0.175)
	aROC	97.678	97.904	98.001	98.907	99.201	99.407	96.248	96.808	**97.199**
		(1.087)	(1.477)	(3.297)	(0.448)	(0.478)	(0.745)	(1.832)	(2.357)	(5.490)
RTD	TPR	72.078	64.870	49.875	83.795	78.020	63.700	28.270	26.220	26.675
		(6.953)	(9.456)	(17.984)	(5.558)	(8.661)	(15.971)	(5.487)	(7.676)	(17.766)
	TNR	96.894	98.150	99.494	98.198	98.842	99.633	92.025	96.116	99.259
		(0.774)	(0.498)	(0.182)	(0.617)	(0.456)	(0.162)	(0.611)	(0.405)	(0.180)
	aROC	96.132	96.306	96.464	98.842	98.966	99.108	62.857	63.507	62.839
		(1.650)	(1.982)	(3.853)	(0.731)	(0.771)	(1.011)	(3.951)	(5.541)	(12.551)
FSD	TPR	74.895	69.595	54.650	86.188	82.805	69.975	29.573	27.100	26.900
		(4.479)	(6.845)	(17.210)	(3.312)	(5.199)	(15.632)	(5.450)	(7.609)	(17.761)
	TNR	97.211	98.400	99.542	98.465	99.095	99.697	92.175	96.163	99.262
		(0.498)	(0.360)	(0.174)	(0.368)	(0.274)	(0.158)	(0.606)	(0.400)	(0.179)
	aROC	96.790	97.133	97.293	99.139	99.356	99.472	60.811	61.425	60.631
		(1.158)	(1.549)	(3.436)	(0.387)	(0.410)	(0.637)	(3.566)	(4.986)	(12.059)
KMD	TPR	79.235	79.460	76.550	75.550	75.205	67.600	39.155	35.345	27.975
		(4.720)	(6.723)	(17.260)	(4.362)	(6.358)	(15.708)	(4.415)	(6.633)	(17.452)
	TNR	97.693	98.919	99.763	97.283	98.695	99.673	93.239	96.597	99.272
		(0.524)	(0.354)	(0.174)	(0.485)	(0.335)	(0.159)	(0.491)	(0.349)	(0.176)
	aROC	97.343	98.080	98.639	97.707	98.703	99.289	52.172	52.948	56.181
		(1.205)	(1.446)	(3.007)	(0.697)	(0.648)	(0.846)	(2.376)	(2.971)	(6.706)
RML-KMD	TPR	87.253	84.795	77.850	88.695	84.225	70.600	39.673	36.475	29.750
		(4.504)	(6.544)	(17.527)	(4.194)	(6.202)	(16.497)	(4.294)	(6.534)	(17.109)
	TNR	98.584	99.200	99.776	98.744	99.170	99.703	93.297	96.657	99.290
		(0.500)	(0.344)	(0.177)	(0.466)	(0.326)	(0.167)	(0.477)	(0.344)	(0.173)
	aROC	**98.662**	**98.732**	98.724	**99.443**	**99.474**	99.472	52.380	52.984	54.898
		(0.935)	(1.272)	(2.932)	(0.361)	(0.416)	(0.680)	(2.173)	(2.796)	(7.098)
GKD	TPR	**88.398**	**86.375**	**79.350**	**90.860**	**87.515**	**74.850**	**86.268**	**82.835**	**67.525**
		(3.949)	(6.101)	(17.540)	(3.372)	(5.669)	(18.001)	(4.033)	(6.118)	(19.529)
	TNR	**98.711**	**99.283**	**99.791**	**98.984**	**99.343**	**99.746**	**98.474**	**99.097**	99.672
		(0.439)	(0.321)	(0.177)	(0.375)	(0.298)	(0.182)	(0.448)	(0.322)	(0.197)
	aROC	98.180	98.583	**98.803**	98.953	99.285	**99.587**	**96.579**	**96.849**	97.142
		(0.917)	(1.217)	(2.831)	(0.415)	(0.378)	(0.478)	(1.710)	(2.234)	(5.408)

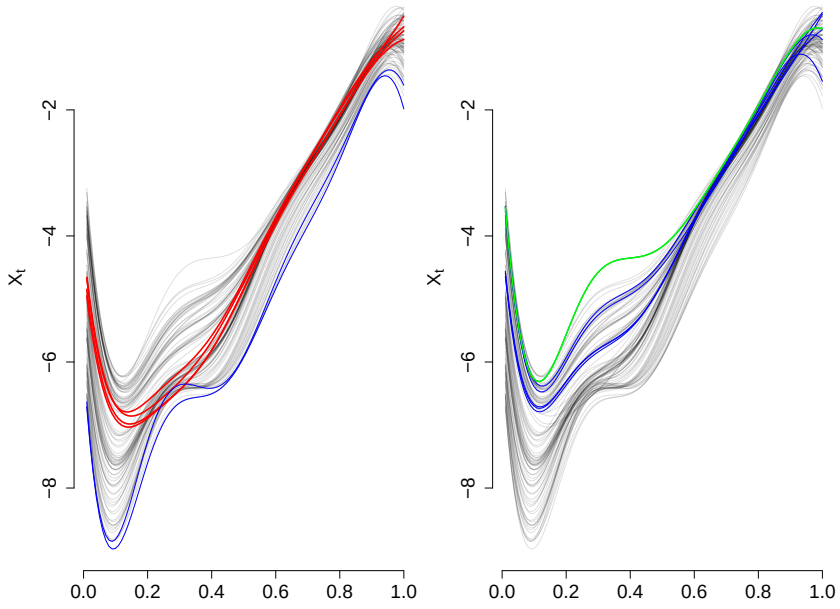


Figure 3.4: Australian Mortality data: regular curves in black ("—") and outliers detected in colours red ("—"), blue ("—") and green ("—") by the GKD (left) and by the KMD (right), for $\nu = 5\%$. In red (—) we have highlighted the curves detected as outliers that belongs to the period 1942–1945, in green (—) the year 1919; remaining outliers in blue (—).

3.4.2 Detecting outliying curves in the Australian mortality rate database

For this experiment we consider age-specific log-mortality rates for Australian males. The source of the data is the Australian Demographic Data Bank which is publicly available in the R-package `'fds'` Shang and Hyndman (2013). In Figure 3.4 each curve correspond to one year from 1901-2003 and is defined for age cohorts from 0 to 100 years. As expected, for low-age cohorts (until 12 years, approximately), the mortality rates present a decreasing trend and then start to grow until late ages, where all cohorts achieve a 100% mortality rate.

After smoothing the data –with the methodology proposed in Chapter 2–, we carried out the outlier identification process. In this experiment we do not know a priori if there is any outlying curves, so after having conducted inference over a wide range of values for ν, we defined as outlier, or abnormal curve, those ones above the 0.95-quantile –i.e. the 5% most distant to the center paths–. Results are illustrated in Figure 3.4.

In a previous work Arribas-Gil and Romo (2014), the authors identified as a "shape" outlier to the curve corresponding to the mortality rate of the year 1919, which corresponds to the influenza pandemic episode that causes around $15,000$ dead as the virus

spread through Australia. As it is shown in Figure 3.4 the KMD is also able to detect this outlier. Moreover the GKD proposed is able to identify as outlier the curves corresponding to years 1942 to 1945, associated to the Second World War, in which Australia participated. Regarding this last, competitor measures, included the methodology proposed by Hyndman (1996), are not able capture these anomalies in its entirely. The curves corresponding to these years present a different shape with respect to the rest of the curves. Likewise, they do not present any extreme point (age-cohort) that could help to infer that their shape is different from the bulk of data. In particular for the cohorts from 15 to 40 years it can be appreciated the difference in the pattern of the curves, so they can be considered as shape outliers.

In Table 3.2, are presented the full results of the anomaly detection exercise considering the KMD, the $RML--KMD$, GKD and the results obtained with other competitor procedures described previously, for $\nu = \{0.1, 0.05, 0.01\}$. As is expected, the use of an inappropriate value for ν increases the number of false positives in the analysis. A convenient criterion for choosing the value of ν is toconsider the ratio:

$$\{KMD(\tilde{x}_s(t), \mathcal{S}_n) / \sum_{s=1}^{n} \text{KMD}(\tilde{x}_s(t), \mathcal{S}_n)\}_{s=i}^{n},$$

where $KMD(\tilde{x}_s(t), \mathcal{S}_n)$ represents the KMD for the curve s. Using this criterion, we have fixed $\nu = 5\%$. As it can be seen in Figure 3.5, the distribution of the estimated robust $RML - KMD$ presents an elbow at point 0.022, and this corresponds to a value of $\nu = 5\%$. The competitor measures are: the modified band depth (MBD) López-Pintado and Romo (2009), the random Tukey depth (RTD) Cuevas et al. (2007), the h-mode depth $h - MD$ Cuesta-Albertos and Nieto-Reyes (2008), the functional spatial depth (FSD) Sguera et al. (2014), the Bagplot Hyndman (1996) and the outliergrams Arribas-Gil and Romo (2014).

When $\nu = 5\%$, most of the competitor measures identify as anomalous curves the years that correspond to the first and last years of the sample, and the influenza pandemic episode in 1919. Even though it is true that for the early 2000s, the mortality rates are the lowest ones, they present the same dynamic as the rest of the years of the sample, so they could be considered as false-positive identifications. The temporal dynamic implicit in the data shows that the mortality rate decreases systematically every year for all the cohorts. This means that a curve that is far from the "center" of the distribution is not necessarily an anomalous curve, but follows the natural dynamics of the process that generates the samples every year. With respect to the proposed kernel depth meth-

Table 3.2: Anomalous years detected by the different methods for different values of ν

Method	**Anomalous years**		
	$\nu = 10\%$	$\nu = 5\%$	$\nu = 1\%$
MBD	1901-1903; 1919; 1997-2003	1901; 1902; 2000-2003	2002; 2003
h-MD	1901; 1902; 1919; 1995; 1997-2003	1919; 1999-2003	1919; 2003
RTD	1901-1903; 1919; 1997-2003	1901; 1902; 1919; 2001-2003	2003
FSD	1901-1904; 1919; 1998- 2003	1901; 1902; 1919; 2001-2003	1919; 2003
KMD	1901; 1904; 1907-1909; 1911; 1914; 1919; 1920; 2002; 2003	1907; 1908; 1914; 1919; 2002; 2003	1919; 2003
RML-KMD	1911-1914; 1916; 1919; 1936-1940	1912; 1914; 1919; 1938-1940	1914; 1919
GKD	1901; 1902; 1941-1945; 1998; 1999; 2003	1942-1945; 1999; 2003	1942; 1943
Adj. Outliergram	1919		
Bagplot	1913; 1914; 1919; 1998-2003		

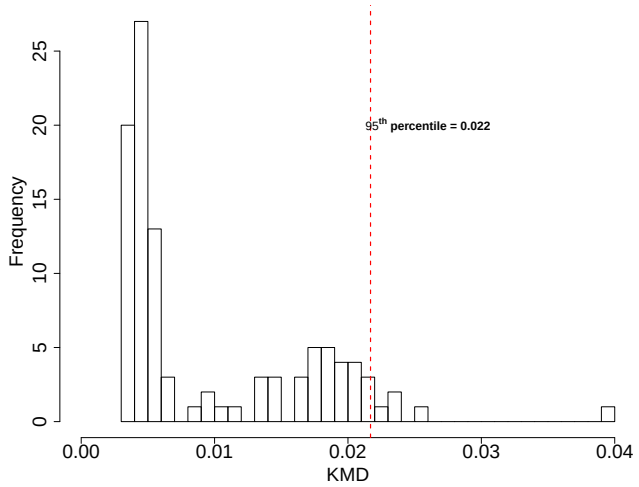


Figure 3.5: Distribution of the $RML - KMD$ for the mortality rate dataset. The vertical red line denotes the 95^{th} percentile of the $RML - KMD$ distribution which corresponds to $\nu = 5\%$.

ods, these are able to identify as anomalous curves year 1919 (influenza pandemic), and those years corresponding to the Second World War, except for the year 1941.

3.4.3 Identifying anomalous numbers

In this experiment we consider as main data a sample of $n = 100$ multivariate raw functional data set $\{(x_1(t), y_1(t)), \ldots, (x_n(t), y_n(t))\}$, where $x(t), y(t)$ represent the 'x' and 'y' coordinates of an image, in this case a number '2' evaluated at $t \in [0, 1]$. Then we contaminate the main sample with ten realizations of the number '3'. In Figure 3.3 we present a sample of the data at hand. The functional approximation of the 'x' and 'y' coordinates is presented in Figure 3.7.

The results presented in table 3.3 show that the GKD and the and the RML-KMD

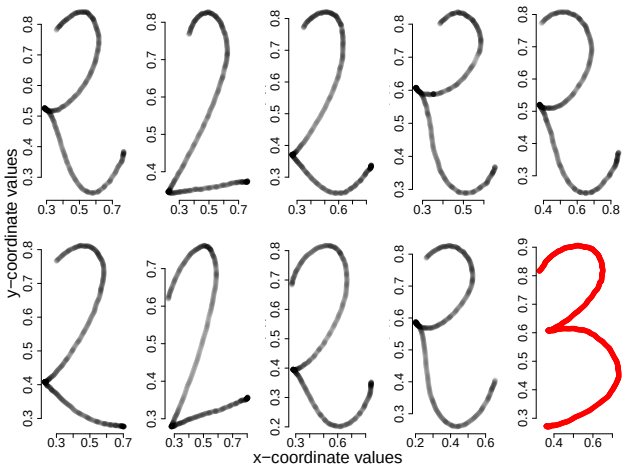


Figure 3.6: Sample of numbers image: regular data ('two's') in black ("—") and outyling filed ('three') in colour red ("—").

proposed are able to capture all the true outliers, as they present a true positive rate of 100% with no false identifications. On the other hand the MMBD and the MFPD detect 70% and 80% of the outliers. The KMD present also good results but they are not able to boost the GKD and RML-KMD results.

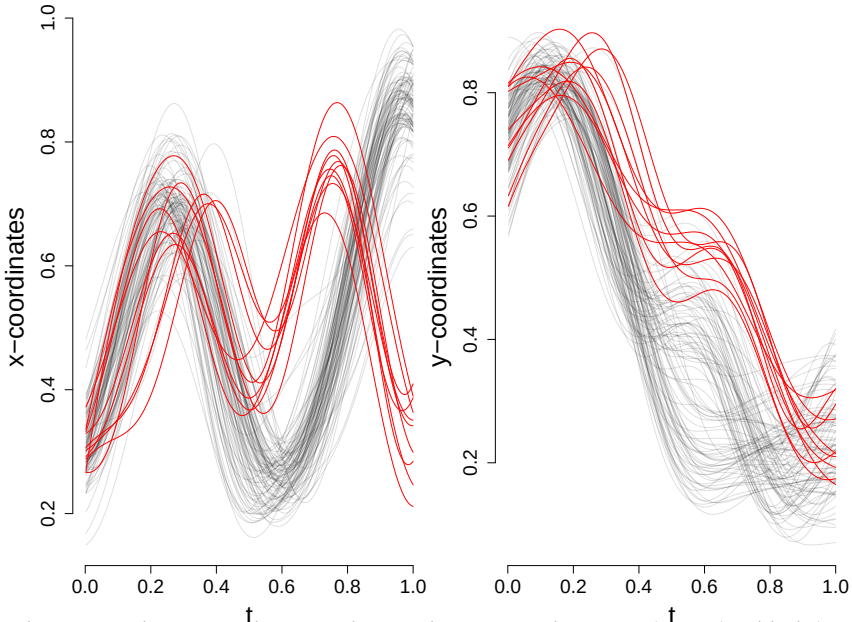


Figure 3.7: Numbers image data: x-coordinates and y-coordinates. Regular curves ('two's') in black ("—") and outliers curves ('three's') in red .

Table 3.3: Sensitivity (TPR), specificity (TNR) and the area under the ROC curves (aROC).

Method	TPR	TNR	aROC
MMBD	0.7	0.97	0.978
MFPD	0.8	0.98	0.993
KMD	0.7	0.97	0.978
RML-KMD	**1**	**1**	**1**
GKD	**1**	**1**	**1**

3.4.4 Identifying anomalous human gestures: a biometric application

In this biometric application we use acceleration data of 4478 human vocabulary gestures samples, recorded from 8 users over an elongated period of time. For each individual an array of $n = 560$ samples with 'x','y' and 'z' gestures coordinates are recorded at $t = 1, \ldots, 315$ points. The experiment that collected the data consisted of 8 participants —2 undergraduate and 8 graduate students, all of them right handed– that recorded 8 vocabulary gestures, 10 times using the wii remote control, see (Liu et al., 2009, Figure 3, p. 5). The gestures were recorded 7 times for each individual within a time window of 3 weeks. If we consider 8 gestures recorded 10 times each, and we repeat 7 times this procedure we obtain the 560 samples for each individual. This research data is publicly available in the R-package `'mfds'`, see Liu et al. (2009).

For this empirical excercise we consider the recorded data by the *fourth* individual as the main data, and contaminate it with $\nu \in \{1\%, 5\%, 10\%\}$ randomly selected samples of invidual *seven* (the 'outlying' data). This procedure was repeated 1000 times so we obtain 1000 random samples contaminated with outliers. To illustrate the data, in Figure 3.8 are shown one instance of the paths for $\nu = 10\%$. For this experiment, the values of the parameter ν are assumed known in each scenario. With respect to the kernel parameter σ, and the penalization parameter of the regularization problem a 10-fold cross-validation procedure was performed using a single sample, which correspond to the first instance. The reference values, which remain fixed throughout the exercise, are $\sigma = 50$ and $\gamma = 0.1^{10}$.

The results presented in Table 3.4 show that the GKD measure proposed outperform the competitor depth measure (MMBD) in the three scenarios considered for the parameter $\nu \in \{1\%, 5\%, 10\%\}$. Comparing among the proposed methods, the GKD and the RML-KMD seems to perform similar, being the main difference with respect to the non-robust estimation of the covariance matrix, that is when we compare with respect to the KMD.

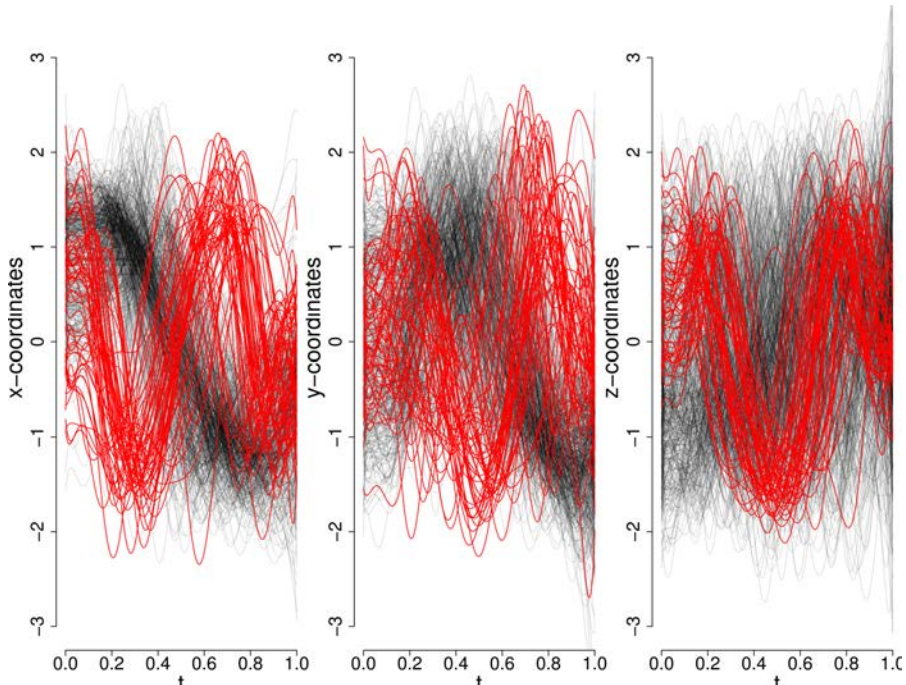


Figure 3.8: 616 Functional data curves, corresponding to the contaminating scenario $\nu = 10\%$. In black ("—"), the sample of regular paths (individual *four*) – 560 –, and abnormal paths (individual *seven*) in red – 56 –.

Table 3.4: Simulation analysis: Contamination percentages ν in rows. In columns, different methods, average sensitivities, specificities and the areas under the ROC curves (aROC) (this last on a scale of 10^2). The corresponding standard-error is reported in parenthesis.

Method	MMBD			KMD			RML-KMD			GKD		
Metric	TPR	TNR	aROC	TPR	TNR	aROC	TPR	TNR	aROC	TPR	TNR	aROC
10%	77.263 (3.874)	96.655 (0.387)	97.212 (0.387)	54.641 (4.142)	94.393 (0.414)	92.081 (1.386)	84.730 (3.638)	97.402 (0.364)	**97.591** (0.907)	**87.936** (3.214)	**97.722** (0.321)	97.139 (0.871)
5%	67.971 (6.411)	98.041 (0.321)	97.090 (0.459)	54.371 (6.074)	97.361 (0.304)	94.845 (1.736)	74.761 (5.322)	98.381 (0.266)	97.543 (1.243)	**82.896** (3.886)	**98.788** (0.194)	**97.805** (1.022)
1%	39.483 (14.978)	99.352 (0.160)	98.343 (0.815)	44.967 (10.583)	99.410 (0.113)	97.264 (3.099)	40.933 (12.852)	99.367 (0.138)	97.609 (2.636)	**68.333** (14.479)	**99.661** (0.155)	**98.429** (2.162)

3.5 Chapter Summary

In this chapter we introduce kernel based depth measures for functional data, i.e. realizations of a stochastic process, represented in a Reproducing Kernel Hilbert Space. Two depth measures that induce order into the data were proposed: i) the Kernel Mahalanbobis Depth (KMD), based on the Mahalanobis distance jointly with a robustified version of it and, ii) the Generalized Kernel Depth (GKD) based on a generalization of the Mahalanobis depth via density kernels. We prove that the proposed Generalized Kernel Depth measure fulfil several desirable theoretical properties, in particular the invariance under RKHS bases choice. Morover we show that the Mahalanobid depth and the h–mode depth are particular cases of the Generalized Kernel Depth proposed.

Simulations results demonstrate that GKD works considerably better than other techniques when the goal is to identify anomalous functional data in non-Gaussian scenarios. Aditionally we conduct an analysis of mortality rate curves as an interesting application in a real–data context –for the univariate functional data case–, where both depth measures show their adequacy to capture anomalous curves, principally associated with the Second World War and the Influenza episode in 1919. With respect to the multivariate functional data excercises both, the GKD and the RML-KMD, show outstanding results obtaining true positive detection rates of 100% in the numbers experiment and almost 90% in the biometric application.

Chapter 4

Entropy measures for stochastic processes

The family of α-Entropies, originally proposed by Rényi et al. (1961), plays an important role in information theory and statistics. Consider a random variable Z distributed according to a measure F that admits a probability density function f. Then for $\alpha \geq 0$ and $\alpha \neq 1$, the α-Entropy of Z is computed as follows

$$H_\alpha(Z) = \frac{1}{1-\alpha} \log\left(V_\alpha(Z)\right) \tag{4.1}$$

where $V_\alpha(Z) = \mathbb{E}_F\{f^{\alpha-1}\}$, and $\mathbb{E}_F$ stands for the expected value with respect to the F measure. Several renowned entropy measures in the statistical literature are particular cases in the family of α-Entropies. For instance, when $\alpha = 0$ we obtain the Hartley entropy, when $\alpha \to 1$ then H_α converges to the Shannon entropy and when $\alpha \to \infty$ then H_α converges to the Min-entropy measure. An interesting question is how to extend this definition when dealing with stochastic processes, and in particular how to estimate the Entropy of a stochastic process with a set of random functions, namely a sample of realizations of a stochastic process.

The contribution of this chapter is twofold. Firstly we propose a natural definition of Entropy for stochastic processes that extends the previous one and a suitable sample estimator for the observation of partial realizations of the process, the typical framework when dealing with functional data. We also show that Minimal Entropy Sets (MES), as formally defined in Section 3, are useful to solve anomaly detection problems, a common task in almost all data analysis context.

The chapter is structured as follows: In Sections 4.1 and 4.1.1 we introduce a definition of Entropy for stochastic process, suitable sample estimators for this measure and a definition of the K–Entropy measure. In Section 4.2 we show how to estimate minimum–entropy sets of a stochastic process in order to discover atypical functional data in a sample. Section 4.3 illustrates the theory with simulations and examples and Section 4.4 concludes the work.

4.1 Entropy of a stochastic process

In this section we extend the definition of Entropy to a stochastic process. For the sequel, and as usual in the case of functional data, let $(\Omega, \mathcal{F}, P)$ be a probability space, where $\mathcal{F}$ is the σ-algebra in Ω and P a σ-finite measure. We consider random elements (functions) $X(\omega, t) : \Omega \times T \to \mathbb{R}$ in a metric space (T, τ). See Chapter 2 for further details. We start by defining the d–truncated Entropy for the process $X(\omega, t)$.

Definition 4.1 (d–truncated Entropy for Stochastic Processes)**.** Let X be a centered stochastic process with continuous covariance function. Consider the truncation $X_d(\omega, t) = \sum_{i=1}^{d} \xi_i(\omega) e_i(t)$ and the random vector $Z = (\xi_1, \ldots, \xi_d)$; then the d–truncated Entropy of X is defined as $H_\alpha(X, d) = H_\alpha(Z)$.

The "approximation error" when computing the Entropy of the stochastic process X with Definition 4.1 decreases monotonically with the number of terms retained in the Karhunen–Loève expansion, at a rate that depends on the decay of the spectrum of the covariance function $K_X(s, t)$. In general, the more autocorrelated the process is, the more quickly converge the eigenvalues of $K_X(s, t)$ to zero. In practical functional data applications, see for instance the mortality–rate curves in Section 4.3, the autocorrelation is usually strong and the truncation parameter d will be small when approximating the entropy of the process. Next example illustrates the definition.

Example 4.1. [Gaussian process] When X is a Gaussian Process (GP), the coefficients in the Karhunen–Loève expansion have the further property that are *independent and zero–mean normally distributed* random variables. Therefore the Shannon Entropy ($\alpha = 1$) of X can be approximated with the truncated version of the GP as follows:

$$H_1(X, d) = \frac{1}{2} \log(2\pi e)^d \det(\mathbf{\Sigma}),$$

where $\mathbf{\Sigma}$ is the diagonal covariance matrix with elements $[\mathbf{\Sigma}]_{i,j} = \mathbb{E}(\xi_i \xi_j)$ for $i, j = 1, \ldots, d$.

• • •

In practice, we can only observe some realizations of the stochastic process X and these observations are sparsely registered. Therefore, to estimate the entropy of $X(\omega, t)$ from a random sample of discrete realizations of a stochastic process, a first task is the representation of these paths by means of continuous functions. To this end, we consider a Reproducing Kernel Hilbert Space $\mathcal{H}$ of functions, associated to a positive definite and symmetric kernel function $K : T \times T \to \mathbb{R}$ –see Chaper 2 for further details–.

4.1.1 Estimating Entropy in a Reproducing Kernel Hilbert Space

Definition 4.2 (K–Entropy estimation of a Stochastic Process)**.** Let $\{x_1(t_i), \ldots, x_n(t_i)\}$ for $i = 1, \ldots, m$, be a discrete random sample of X, and let $\{(\lambda_j, \mathbf{v}_j)\}_{j=1}^{d}$ be the eigen–pairs of the kernel matrix $\mathbf{K} \in \mathbb{R}^{m \times m}$, where $d = \operatorname{rank}(\mathbf{K})$. Consider the corresponding finite dimensional representation $S_n := \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ where $\mathbf{z}_l = (\widehat{\xi}_{l,1}, \ldots, \widehat{\xi}_{l,d}) \in \mathbb{R}^d$ for $l = 1, \ldots, n$, and $\widehat{\xi}_{l,j} = \sqrt{\lambda_{l,j}} \sum_{i=1}^{m} a_{l,i} v_{i,j}$ for $j = 1, \ldots, d$. Then, the estimated Kernel Entropy of X is defined as $\widehat{H}_\alpha(X, K) = \widehat{H}_\alpha(Z)$.

In Definition 4.2, $\widehat{H}_\alpha(Z)$ denotes the estimated entropy using the –finite dimensional– representation coefficients $S_n = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$. In Section 4.2 we formally introduce two approaches to estimate Entropy departing from S_n. Next example illustrates the estimation procedure in the context of GPs in Example 4.1.

Illustration with Example 1: Consider 100 realizations of a GP as follows: 50 curves from $X(t) = \sum_{i=1}^{3} \xi_i e_i(t)$ and another 50 curves from $Y(t) = \sum_{i=1}^{3} \zeta_i e_i(t)$; where $e_i(t)$ is a Fourier basis in $T = [0, 1]$, $\xi_i \sim N(\mu = 0, \sigma^2 = 0.5)$, and $\zeta_i \sim N(\mu = 0, \sigma^2 = 2)$ are independent normally distributed r.v. for $i = 1, 2, 3$.

In Figure 4.1 (left) we illustrate the realizations of the stochastic processes, in black ("—") the sample paths of $X(t)$ and in red ("—") the paths corresponding to $Y(t)$. In Figure 4.1 (right) we show the distribution of the linear combination coefficients $\{(\widehat{\xi}_1, \widehat{\xi}_2, \widehat{\xi}_3)_l, (\widehat{\zeta}_1, \widehat{\zeta}_2, \widehat{\zeta}_3)_l\}_{l=1}^{50}$ corresponding to these paths. Following Example 4.1, we estimate the covariance functions $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\xi}}$ and $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\zeta}}$ using the respective coefficients, and plug-in this covariance matrix into the Shannon Entropy expression to obtain the estimated entropies $\widehat{H}_1(X) = 1.402$ and $\widehat{H}_2(Y) = 99.552$, similar to the true entropies $H_1(X) = 1.428$ and $H_2(Y) = 91.420$ respectively. A formal estimation procedure is proposed in Algorithm 2.

An interesting excercise is to analyze the convergence of the estimated entropy to the true values. To this end a Monte Carlo study is presented considerin the Gaussian

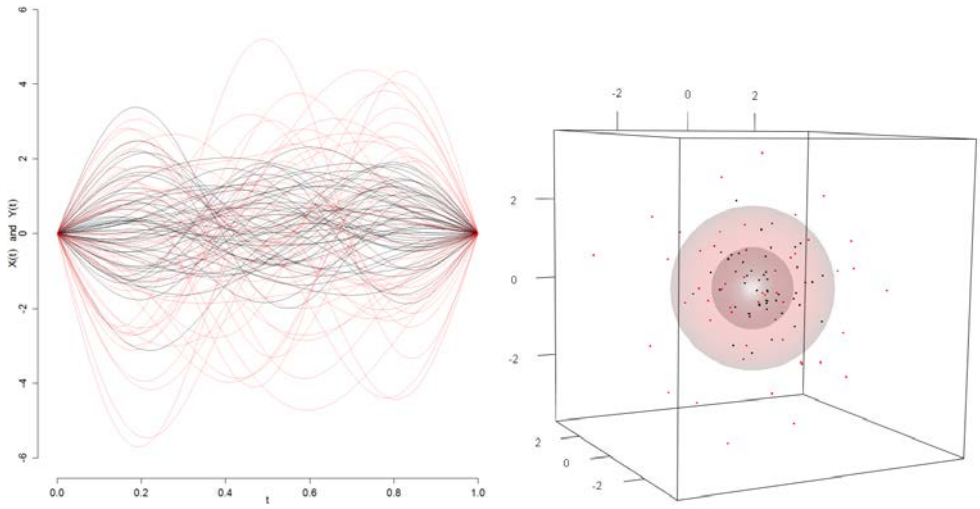


Figure 4.1: Gaussian procesess realizations on the left and coeficients for Entropy estimation on the right. The sizes of the balls on the right are proportional to the determinants of $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\xi}}$ (in black) and $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\zeta}}$ (in red).

processes of Example 4.1. Recalling the processes $X(t)$ and $Y(t)$, the Shannon entropies ($\alpha = 1$) are:

$$H_1(X, d = 3) = \frac{1}{2}\log(2\pi e)^3 \det(\mathbf{\Sigma}_X) \text{ and } H_1(Y, d = 3) = \frac{1}{2}\log(2\pi e)^3 \det(\mathbf{\Sigma}_Y),$$

where $\mathbf{\Sigma}_X$ and $\mathbf{\Sigma}_Y$ are the respective covariance matrices. The Monte Carlo experiment was carried out as an assessment of the uncertainty of the Entropy estimation as the sample size is increased. For the numerical excercies were considered $M = 10.000$ samples from the distribution of ξ_i and ζ_i, for different sample size $N = \{5, 10, 20, 50, 75, 100, 200, 250, 350, 500, 750, 1000, 1500, 2000, 3000\}$. The results, illustrated in Figure 4.2, show that in both cases the estimated Entropy converges relatively fast to the true values, $H_1(X) = 1.428$ and $H_2(Y) = 91.420$.

The choice of kernel parameters in Algorithm 2 is made by cross–validation. This ensures that the curve fitting method is asymptotically optimal. Nonetheless, although the selection of the kernel parameters affects the scale of the estimated Entropy, the center-outward ordering induced by $H_\alpha(X, K)$, as formally proposed in next section, is unaffected. In Section 4.2, we present relevant experimental results to illustrate this property, which make the method robust in terms of the selection of the kernel and regularization parameters.

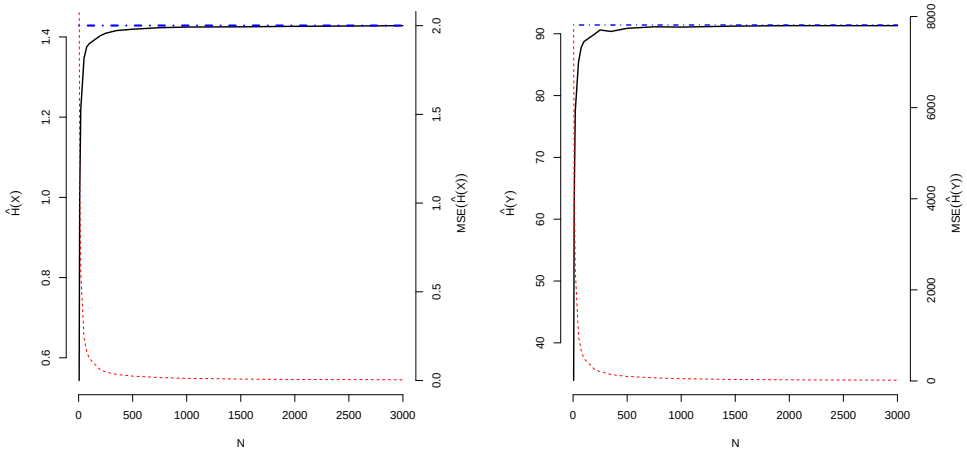


Figure 4.2: Entropy estimation in black (—) , Entropy true value in blue (- · -) and Mean Squared Error in red (- -) for the two Gaussian processes $X(t)$ (left) and $Y(t)$ (right).

Algorithm 2: Estimation of $H_\alpha(X, K)$ from a sample of random paths.

Functional K-Entropy: $(\mathbf{X}, K, \alpha, \gamma, d, density)$;

Input : The raw–data matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ –paths in rows–, the kernel function K, the entropy parameter α, the regularization parameter γ, the truncation parameter $d \leq \text{rank}(\mathbf{K})$, and a predefined *density* estimation procedure.

Output: $\widehat{H}_\alpha(X, K)$

for *l in 1 to n* **do**
  compute $\boldsymbol{a}_l = (\gamma m \mathbf{I} + \mathbf{K})^{-1} \mathbf{y}_l$;
  for *j in 1 to d* **do**
    $\widehat{\xi}_{l,j} = \sqrt{\lambda_j} \sum_{i=1}^{m} a_{l,i} v_{i,j}$
  end
  store $\mathbf{z}_l = (\widehat{\xi}_{1,l}, \ldots, \widehat{\xi}_{d,l})$
end
Consider $S_n = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ an iid sample from the random vector Z. Estimate $\widehat{F}_Z$ with a predefined *density estimation* procedure and compute $\widehat{V}_\alpha(Z) = \mathbb{E}_{\widehat{F}}\{f^{\alpha-1}\}$;
Return $\widehat{H}_\alpha(X, K) = \widehat{H}_\alpha(Z)$.

4.2 Minimum Entropy for anomaly detection

Anomaly detection is a common task in almost all data analysis context. The unsupervised approach considers a sample $X_1, \ldots, X_n$ of random elements where most instances follow a well defined pattern and a small proportion, here denoted as $\nu \in [0, 1]$, present an abnormal pattern. In recent works, see for instance López-Pintado and Romo (2009); Cuevas et al. (2007); Cuesta-Albertos and Nieto-Reyes (2008); Chakraborty and Chaudhuri (2014b), the authors propose depth measures and related methods, to deal

with functional outliers. In this section we propose a novel criterion to tackle the problem of anomaly detection with functional data using the ideas and concepts developed in Section 4.1.1. For a real–valued d–dimensional random vector Z that admits a continuous density function f_Z,define $H_\alpha(A_Z) = \frac{1}{1-\alpha}\log\left(\int_A f_Z^\alpha(\mathbf{z})d\mathbf{z}\right)$ to be the entropy of the Borel–set A with respect to the measure F_Z. Then, the ν–minimum–entropy set (MES) is formally defined as

$$\text{MES}_\nu(Z) := \{\arg\min_{A\subset\mathbb{R}^d} H_\alpha(A_Z) \text{ s.t. } P(A) \geq 1-\nu\}.$$

The MES_ν is equivalent Hero (2007); Xie et al. (2016) to a ν-high density set (HDS) Hyndman (1996) formally defined as $\text{HDS}_\nu(Z) = \{\mathbf{z} \in \mathbb{R}^d | \, f_Z(\mathbf{z}) > c_\nu\}$, where c_ν is the largest constant such that $P(\text{HDS}_\nu(Z)) \geq 1-\nu$, for $0 < \nu < 1$. Therefore the complement of MES is a suitable set to define outlier data in the sample, considering $\tilde{x}(t) \notin \text{MES}_\nu$ as an atypical realization of X. Next we give two approaches to estimate MES.

4.2.1 Parametric approach

Given a random sample of n discrete random paths $\{x_1(t_i), \ldots, x_n(t_i)\}$ for $i = 1, \ldots, m$, we transform this sample into d–dimensional vectors $S_n = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ using the representation and truncation method proposed in this work, numerically implemented in lines 2–to–8 in Algorithm 2. Assume further that $f_Z(\mathbf{z}, \boldsymbol{\theta})$ is a suitable probability model for the random sample $\mathbf{z}_1, \ldots, \mathbf{z}_n$, then we estimate by robust maximum likelihood (RML) the parameters $\boldsymbol{\theta}$. For instance, in this chapter we consider $f_Z(\mathbf{z}, \boldsymbol{\theta})$ to be the normal density, and then RML estimated parameters are $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$, the robust mean vector and covariance matrix respectively.

For details on robust estimation, we refer to Maronna et al. (2018). After the estimation of the distribution parameters, the computation of H_α follows by plug–in the estimated density $f_Z(\mathbf{z}, \widehat{\boldsymbol{\theta}})$ in Equation 4.1. Moreover, for the normal model, the estimated set MES_ν is defined trough the following expression

$$\text{MES}_\nu(S_n) = \{\mathbf{z} \in \mathbb{R}^d | (\mathbf{z} - \widehat{\boldsymbol{\mu}})^T \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{z} - \widehat{\boldsymbol{\mu}}) \leq \chi^2_d(\nu)\},$$

where $\chi^2_d(\nu)$ is the $1-\nu$ quantile of a Chi–square distribution with d–degrees of freedom. Then if the coefficient $\mathbf{z}_i$, representing $\tilde{x}_i(t)$, lies outside this ellipsoid; we say that the functional datum is atypical. When the proportion of outlier ν in the sample is known a–priori, the $\chi^2_d(\nu)$–quantile can be replaced by the corresponding sample $1-\nu$ Mahalanobis distance quantile, as is the case in the Section 4.3.1.

4.2.2 Non–parametric approach

The following are definitions to introduce further non–parametric estimation methods. For the random vector $Z \in \mathbb{R}^d$ distributed according to F_Z, let $B_Z(\mathbf{z}, r_\delta) \subset \mathbb{R}^d$ be the $\mathbf{z}$–centered ball with radius r_δ that fulfills the condition $\delta = \int_{B_Z(\mathbf{z},r_\delta)} f_Z(\mathbf{z})d\mathbf{z}$, then the δ**-Neighbors** of the point $\mathbf{z}$ is the open set $\Delta_\mathbf{z} = \mathbb{R}^d \bigcap B(\mathbf{z}, r_\delta)$.

Definition 4.3 (δ**-Local** α**-Entropy).** Let $\mathbf{z} \in \mathbb{R}^d$, for $\alpha > 0$ and $\alpha \neq 1$, the δ-local α-entropy of the r.v. Z is

$$h_\alpha(\Delta_\mathbf{z}) = \frac{1}{1-\alpha} \log \left(\int_{\Delta_\mathbf{z}} f_Z^\alpha(\mathbf{z}) d\mathbf{z} \right) \text{ for all } \mathbf{z} \in \mathbb{R}^d.$$

Under mild regularity conditions on f_Z, the local entropy measure is a suitable metric to characterize the degree of abnormality of every point $\mathbf{z}$ in the support of F_Z. Several natural estimators of local entropy measures can be considered, for instance the (average) distance from the point $\mathbf{z}$ to its k^{th}-nearest neighbor. We estimate MES combining the estimated δ-Local α-Entropy. As in the parametric case, let $\{x_1(t_i), \ldots, x_n(t_i)\}$ for $i = 1, \ldots, m$, be a random sample of n discrete random paths, we transform this sample into d–dimensional vectors $S_n = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ following the lines 2–to–8 in Algorithm 2. Next, we estimate the local entropy for this data using the estimator $\widehat{h}_\alpha(\Delta_{\mathbf{z}_i}) = \exp(\bar{d}_k(\mathbf{z}_i, S_n))$, where $\bar{d}_k(\mathbf{z}_i, S_n)$ is the average distance from $\mathbf{z}_i$ to its k^{th}-nearest neighbour Beirlant et al. (1997), and then estimate the MES$_\nu$ solving the following optimization problem

$$\max_{\rho, \epsilon_1, \ldots, \epsilon_n} (1-\nu)\rho - \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \quad \text{s.t. } \hat{h}_\alpha(\Delta_{\mathbf{z}_i}) \geq \rho - \epsilon_i,\ \epsilon_i \geq 0 \text{ for } i = 1, \ldots, n. \tag{4.2}$$

The solution to this problem, ρ^*, leads to the following decision function

$$D(\mathbf{z}) = \text{sign}(\rho^* - \hat{h}_\alpha(\Delta_\mathbf{z})),$$

where $D(\mathbf{z}) = +1$ if $\mathbf{z}$ corresponds to the $(1-\nu)$ proportion of curves projected near the origin, that is, the set of curves that belongs to a low entropy (high density) set. The following theorem shows that as the number of available curves increases, the estimation method asymptotically detect the proportion $1-\nu$ of curves belonging to the MES$_\nu$.

Theorem 4.1. *At the solution of the optimization problem stated in Equation 4.2, the following equality holds*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{z}_i) = 1 - \nu,$$

where $I(\mathbf{z}) = 1$ *if* $\hat{h}_\alpha(\Delta_\mathbf{z}) \leq \rho^*$ *and* $I(\mathbf{z}) = 0$ *otherwise.*

4.3 Experimental section

The aim of this section is to illustrate the performance of the proposed methodology to detect abnormal observations in a sample of functional data. In what follows, for the representation of functional data, we consider the Gaussian kernel function $K(t_l, t_k) = e^{-\sigma\|t_l - t_k\|^2}$. The kernel parameter σ and the regularization coefficient γ in Algorithm 2 were defined through cross-validation.

4.3.1 Simulation analysis

In a Monte Carlo study we investigate the performance of the proposed method over three data configurations (scenarios A, B and C). Specifically, we consider the following generating processes, a fraction $1 - \nu$ of $n = 400$ curves are realizations of the following stochastic model

$$X_l(t) = \sum_{j=1}^{4} \xi_j \sin(j\pi t) + \varepsilon_l(t), \text{ for } l = 1, \ldots, (1-\nu)n, \text{ and } t \in [0, 1],$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_4)$ is a normally distributed multivariate random variable with mean $\boldsymbol{\mu}_\xi = (4, 2, 4, 1)$ and diagonal co–variance matrix $\boldsymbol{\Sigma}_\xi = \text{diag}(5, 2, 2, 1)$, and $\varepsilon_l(t)$ are independent autocorrelated random error functions.

The remaining proportion of data $n\nu$ with $\nu \in \{1\%, 5\%, 10\%\}$ are outliers that contaminate the sample according to the following typical scenarios (see Cano et al. (2015)):

(A) *Magnitude outliers*: $Y_l(t) = \sum_{j=1}^{4} \zeta_j \sin(j\pi t) + \varepsilon_l(t)$, for $l = 1, \ldots, \nu n$, and $t \in [0, 1]$, where $\boldsymbol{\zeta}$ is a normally distributed multivariate r.v. with parameters $\boldsymbol{\mu}_\zeta = 2.5\boldsymbol{\mu}_\xi$ and $\boldsymbol{\Sigma}_\zeta = (2.5)^2\boldsymbol{\Sigma}_\xi$.

(B) *Shape outliers*: $Y_l(t) = \sum_{j=1}^{4} \zeta_j \sin(j\pi t) + \varepsilon_l(t)$, for $l = 1, \ldots, \nu n$, and $t \in [0, 1]$, where $\boldsymbol{\zeta}$ is a normally distributed multivariate r.v. with parameters $\boldsymbol{\mu}_\zeta = (4, -2, 1, 3)$ and $\boldsymbol{\Sigma}_\zeta = \boldsymbol{\Sigma}_\xi$.

(C) A *combination* considering $\nu n/2$ outliers from scenario A and $\nu n/2$ outliers from scenario B.

To illustrate the generating process, in Figure 4.2 we show one instance of the simulated paths in scenario (C) with $\nu = 10\%$. We test our parametric Entropy (PA)

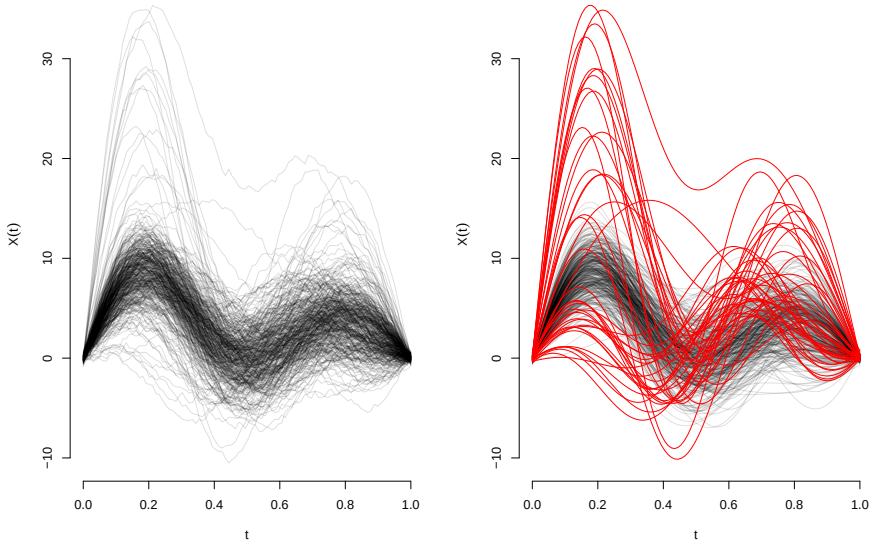


Figure 4.3: Left: Raw data, 400 curves corresponding to scenario C with $\nu = 10\%$. Right: Functional data, in black ("—") the sample of regular paths $X(t)$ and abnormal curves $Y(t)$ in red ("—").

and non–parametric Entropy (NPA) method –already implemented in the R–packages `bigdatadist` Martos and Hernández (2018)– against several well known depth measures for functional anomaly detection, namely: The modified band depth (MBD), the h-mode depth (HMD), the random Tukey depth (RTD), and Functional spatial depth (FSD), see López-Pintado and Romo (2009); Cuevas et al. (2007); Cuesta-Albertos and Nieto-Reyes (2008); Chakraborty and Chaudhuri (2014b) respectively, already implemented in the R–package `fda--usc` Febrero-Bande and Oviedo de la Fuente (2013).

Let P and N be the amount of outlier and normal data in the sample respectively and let TP=True Positive, FP =False Positive, TN = True Negative and FN = False Negative be the respective quantities detected by different methods; in Table 4.1 we report the following average metrics TPR = TP/P (true positive rate or sensitivity), TNR = TN/N (true negative rate or specificity) and PPV = TP / (FP+TP) (precision) of each method obtained through the $M = 1000$ replications in the Monte Carlo study.

As can be seen, the parametric (PA) and non-parametric (NPA) Entropy methods proposed in this article outperform other recently proposed depth measures in the 3 scenarios considered in the experiments.

The parametric approach seems to be slightly (but consistently) more effective than

Table 4.1: Monte–carlo study: Scenarios and contamination percentages ν in columns. In rows, different methods and average sensitivities, specificities and precisions (standard-error reported in parenthesis).

		Scenario A			Scenario B			Scenario C		
Method	Metric	10%	5%	1%	10%	5%	1%	10%	5%	1%
MBD	TPR	74.867	71.010	55.300	48.275	39.395	13.475	67.787	58.365	36.300
		(4.699)	(7.712)	(20.852)	(5.914)	(9.013)	(16.180)	(5.351)	(7.772)	(18.341)
	TNR	97.207	98.474	99.548	94.252	96.810	99.126	96.420	97.808	99.356
		(0.522)	(0.406)	(0.210)	(0.657)	(0.474)	(0.163)	(0.594)	(0.409)	(0.185)
	PPV	7.878	3.653	0.557	5.376	2.092	0.136	7.239	3.041	0.366
		(0.417)	(0.368)	(0.208)	(0.589)	(0.460)	(0.164)	(0.490)	(0.381)	(0.184)
HMD	TPR	92.665	91.545	88.675	66.532	62.780	47.475	79.992	76.765	66.025
		(3.295)	(5.173)	(14.793)	(6.084)	(8.809)	(21.206)	(4.562)	(7.039)	(18.004)
	TNR	99.185	99.555	99.885	96.281	98.041	99.469	97.776	98.777	99.656
		(0.366)	(0.272)	(0.149)	(0.676)	(0.463)	(0.214)	(0.506)	(0.370)	(0.181)
	PPV	9.402	4.615	0.888	7.124	3.256	0.478	8.328	3.927	0.664
		(0.272)	(0.237)	(0.146)	(0.559)	(0.428)	(0.201)	(0.396)	(0.332)	(0.171)
RTD	TPR	83.555	83.045	76.400	50.972	43.940	22.700	71.975	65.225	49.700
		(4.743)	(0.694)	(18.931)	(9.409)	(1.279)	(2.1334)	(7.178)	(9.716)	(1.834)
	TNR	98.174	99.104	99.762	94.544	97.049	99.218	96.889	98.165	99.491
		(0.526)	(0.365)	(0.191)	(1.045)	(0.674)	(0.215)	(0.798)	(0.511)	(0.184)
	PPV	8.633	4.220	0.766	5.629	2.317	0.229	7.613	3.373	0.501
		(0.406)	(0.323)	(0.187)	(0.930)	(0.648)	(0.215)	(0.648)	(0.468)	(0.183)
FSD	TPR	81.472	83.215	81.925	50.275	46.550	27.400	74.775	69.485	53.775
		(3.978)	(5.947)	(16.671)	(5.238)	(8.018)	(19.547)	(4.601)	(6.859)	(16.707)
	TNR	97.941	99.116	99.817	94.475	97.186	99.267	97.197	98.396	99.533
		(0.442)	(0.313)	(0.168)	(0.582)	(0.421)	(0.197)	(0.511)	(0.361)	(0.168)
	PPV	8.457	4.230	0.821	5.576	2.455	0.277	7.870	3.581	0.542
		(0.344)	(0.277)	(0.164)	(0.516)	(0.402)	(0.197)	(0.409)	(0.328)	(0.166)
Entropy–PA	TPR	**94.150**	**93.215**	**91.725**	**80.740**	**77.390**	66.925	**87.550**	84.935	77.650
		(3.078)	(4.817)	(12.591)	(6.250)	(8.550)	(20.330)	(4.632)	(6.604)	(17.015)
	TNR	**99.350**	**99.649**	**99.916**	**97.860**	**98.810**	99.664	**98.616**	99.207	99.774
		(0.342)	(0.253)	(0.127)	(0.694)	(0.450)	(0.205)	(0.514)	(0.347)	(0.171)
	PPV	**9.524**	**4.691**	**0.918**	**8.390**	**3.955**	0.770	**8.974**	4.349	0.804
		(0.252)	(0.220)	(0.124)	(0.544)	(0.404)	(0.202)	(0.391)	(0.307)	(0.168)
Entropy–NPA	TPR	92.725	91.505	89.050	74.215	77.145	**71.250**	87.225	**85.805**	**79.775**
		(3.325)	(5.228)	(14.630)	(6.237)	(7.904)	(19.970)	(4.217)	(6.198)	(16.788)
	TNR	99.191	99.552	99.889	97.135	98.792	**99.709**	98.586	**99.252**	**99.795**
		(0.369)	(0.275)	(0.147)	(0.693)	(0.416)	(0.201)	(0.468)	(0.326)	(0.169)
	PPV	9.407	4.613	0.892	7.817	3.944	**0.775**	8.948	**4.350**	**0.810**
		(0.274)	(0.240)	(0.144)	(0.557)	(0.373)	(0.198)	(0.356)	(0.287)	(0.166)

the non parametric approach in Scenario A. For Scenarios B and C both methods provide similar results. It is important to remark that the PA method is specially adequated for Gaussian data, while the NPA method does not assume any distributional hypothesis on the data. The simulation results show the robustness of the non parametric approach even when competing with parametric methods designed for specific distributions.

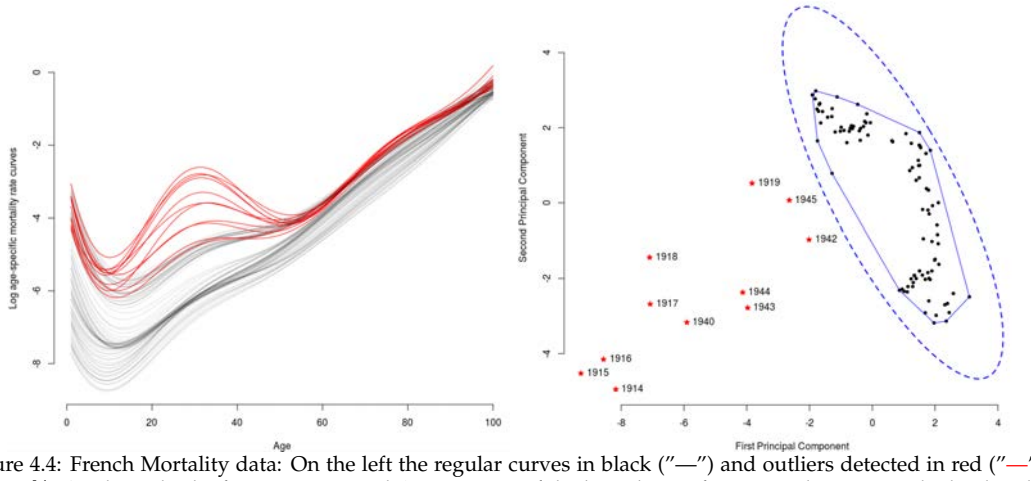


Figure 4.4: French Mortality data: On the left the regular curves in black ("—") and outliers detected in red ("—") for $\nu = 10\%$. On the right the first two Principal Components of the kernel eigenfunctions, the area inside the doted blue ellipsoid (- -) correspond PA estimation of $\text{MES}_{\nu=90\%}$ and the region inside the convex hull in blue (—) to the NPA estimation. The regular curves, represented with black dots (•), lies inside the $\text{MES}_{\nu=90\%}$ and detected outliers with red asterisk ($*$) outside of $\text{MES}_{\nu=90\%}$.

4.3.2 Outliers in the context of mortality–rate curve analysis

We consider the French mortality rates database, available in the R–package `Demography` Hyndman (2017), to study age–specific male death rates in logarithmic scale. In Figure 4.3 (left) each curve corresponds to one year from 1901 to 2006 –106 paths in total– and accounts for the number of deaths per $1,000$ of the mean population in the age group (from 0 to 101 years) in question. As expected, for low–age cohorts (until 12 years approximately), the mortality rates present a decreasing trend and then start to grow until late ages where all cohorts achieve the 100% mortality rate.

For some years the evolution pattern of mortality presents an atypical behavior, mostly coinciding with the first and second World Wars, jointly with the influenza pandemic episode that took place in 1919.

In this experiment we do not know a priori the proportion of atypical curves. Therefore after having conducted inference over a wide range of values for ν, as a way to assess the sensitivity and reliability of the inference when determining the number of abnormal curves, we decided to fix $\nu = 10\%$. In Figure 4.3 (left) we highlight in red the anomalous detected curves with both the Entropy–PA and NPA methods corresponding to the years 1914–to–1919 and $1940, 1942$–to–1945, that match with men (between 20 and 40 years old) participating in I and II World Wars.

In Figure 4.3 (right) we use the first two Principal Components of the kernel eigen-

functions to project the representation coefficients (in this experiment in $\mathbb{R}^{14}$) in two dimensions. As can be seen, the points laying outside the $\text{MES}_{\nu=90\%}$, represented with a doted–blue ellipses when estimating it with PA (- -) and the convex hull with continues–blue line (—) when estimating it with NPA, correspond to the the atypical curves in the sample.

In Tables 4.2 and 4.3, we present the full results of the anomaly detection exercise considering entropy-PA and entropy-NPA and the results obtained with other measures described in Section 4.3 for $\nu = \{0.5, 0.25, 0.15, 0.1, 0.05, 0.01\}$. In the first three scenarios, that is when $\nu = \{0.5, 0.25, 0.15\}$, the results for the competitor measures show that only the HMD is able to capture almost all curves corresponding to the First and Second World War (except year 1941) and the influenza pandemic for a value of $\nu = 0.25$. As is expected, the use of an inappropriate value for ν increases the number of false positives in the analysis. A convenient criterion for choosing the value of ν is to consider the ratio

$$D_M(\mathbf{z}_{[i]}, \hat{\mu_z}) / \sum_{i=1}^{n} D_M(\mathbf{z}_{[i]}, \hat{\mu_z}),$$

where $D_M(\mathbf{z}_{[i]}, \hat{\mu_z})$ represents the Mahalanobis distance sorted in deceasing order of the vector $\mathbf{z}_{[i]}$ representing a curve in the sample (in the case of non-parametric approach, we consider the sorted sequence of estimated local entropies). Using this criterion, in Section 4.2, we decided to fix $\nu = 0.1$, since, as can be seen in Figure 4.5, the distributions of the estimated robust Mahalanobis distances (left) and the local entropies (right) show an elbow at points 10 and 4 respectively, and this corresponds to a value of $\nu = 0.1$ in both cases.

As can be appreciated in Tables 4.2 and 4.3, when $\nu = 0.1$, most of the competitor measures identify as anomalous curves the years that correspond to the First World War and the last years of the sample. Only the HMD is able to partially identify as outliers some years corresponding to the Second World War. Even though it is true that for the early 2000s, the mortality rates are the lowest ones, they present the same dynamic as the rest of the years of the sample, so they could be considered as false-positive identifications. The temporal dynamic implicit in the data shows that the mortality rate decreases systematically every year for all the cohorts. This means that a curve that is far from the "center" of the distribution is not necessarily an anomalous curve, but follows the natural dynamics of the process that generates the samples every year.

With respect to the proposed entropy methods, these are able to identify as anoma-

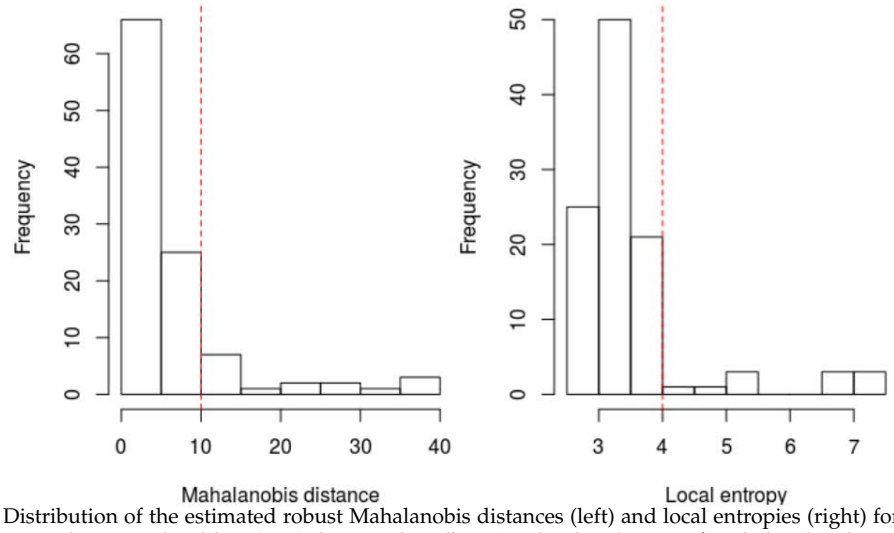


Figure 4.5: Distribution of the estimated robust Mahalanobis distances (left) and local entropies (right) for the mortality rate dataset. The vertical red line (- - -) denotes the 'elbow' in the distribution of Mahalanobis distance and local entropies, respectively, and corresponds to $\nu = 0.1$ in both cases.

lous curves those years corresponding to the First and Second World War, except for the year 1941. Additionally, the entropy methods are the only ones capable of identifying the year 1919 (influenza pandemic) as an outlying curve. Last, but not least, it is important to mention that for the NPA, the obtained results are robust with respect to the number of neighbors k considered in the method.

4.3.3 On order Invariance Property and Robustness

The entropy measure of a stochastic process is a 'K-entropy', which means that the estimated entropy depends on the choice of a particular kernel. In this sense, is the order in the sampled curves –from most to least depth curves– induced by the entropy measure invariant to changes in the kernel function? In this section what we numerically show, is that the order induced by the entropy does not depend on the the kernel function –or its parameters– when representing the functional data at hand. To illustrate this, we constructed an experiment considering Scenario A in Section 4.3 of this chapter when $n = 1000$ and $\nu = 0.05$. As the aim of this section is to show the order invariance property, we consider two different kernel function and different parameters, namely:

i) The Gaussian kernel function

$$K_G(t_l, t_k) = e^{-\sigma \|t_l - t_k\|^2}, \text{with } \sigma = 5, 10, 15.$$

Table 4.2: Anomallus years detected by the different methods for different values of ν.

	Anomalous years		
Metric	50%	25%	15%
MBD	1900-1919; 1922; 1925-1926; 1929; 1940-1944; 1982-2006	1900-1901; 1905-1907; 1909; 1911; 1914-1918; 1940; 1944; 1994-2006	1900; 1907; 1914-1915; 1917-1918; 1940; 1998-2006
HMD	1900-1907; 1914-1919; 1934-1954; 1956; 1989-2006	1900; 1914-1919; 1939-1940; 1943-1951; 1998-2006	1914-1919; 1940; 1943-1944; 1946-1948; 2003-2006
RTD	1900-1921; 1925; 1929; 1940; 1943-1945; 1981-2006	1900-1907; 1911; 1914;1919; 1940; 1944; 1996-2006	1900-1901; 1914-1919; 1944; 1998; 2000; 2002-2006
FSD	1900-1921; 1925-1926; 1940; 1943-1945; 1981-2006	1900-1907; 1914-1919; 1940; 1944; 1995-2006	1900; 1914-1918; 1944; 1998-2006
Entropy-PA	**1901; 1904; 1906; 1912; 1914-1922; 1925; 1931-1932; 1934; 1940-1951; 1954; 1959; 1969; 1986-2006**	**1914-1919; 1940; 1942-1945; 1991-2006**	**1914-1919; 1925; 1934; 1940-1945; 2004; 2006**
Entropy-NPA	**1900-1902; 1911; 1914-1919; 1925-1926; 1931; 1940-1945; 1949; 1955; 1957; 1958; 1961-1965; 196-1982; 1988-1995; 1999; 2004-2006**	**1901; 1914-1919 1931; 1940-1945; 1958; 1970-1971; 1974-1975; 1977-1979; 1990-1993; 2006**	**1914-1919; 1931; 1940; 1942-1945; 1970; 1975; 1978; 1992**

Table 4.3: Anomallus years detected by the different methods for different values of ν.

	Anomalous years		
Metric	10%	5%	1%
MBD	1900; 1915; 1918; 1940; 2000-2006	1900; 2002-2006	2004;2006
HMD	1914-1918; 1940; 1943; 1944; 1946; 1947; 2006	1914-1918; 1940; 1944;	1918; 1940
RTD	1900; 1914-1918; 2002-2006	1915; 1918; 2003-2006	2005;2006
FSD	1914-1918; 2002-2006	1914; 1915; 1918; 2004-2006	1915; 2006
Entropy-PA	**1914-1919; 1940; 1942-1945**	**1914-1918;1940**	**1914; 1915**
Entropy-NPA	**1914-1919; 1940; 1942-1945**	**1914-1918;1940**	**1914; 1915**

The neighbors considered for the NPA was 50.

ii) The Spline kernel function

$$K_S(t_l, t_k) = \prod_{d=1}^{D} 1 + t_l t_k + t_l t_k \min(t_l, t_k) - \frac{t_l + t_k}{2} \min(t_l, t_k)^2 + \frac{t_l + t_k}{3} min(t_l, t_k)^3.$$

The results, displayed in Figure 4.6 in the case of the parametric approach (left panel) and for the non–parametric approach (right panel), show that the order induced in the sample curves by the entropy measure is invariant to changes in the kernel function considered. This property makes the method robust in terms of the selection of the kernel and regularization parameters. This exercise was also carried out for different sample sizes, $n = \{2000, 3000, 5000\}$ and different values of parameter ν, with similar results.

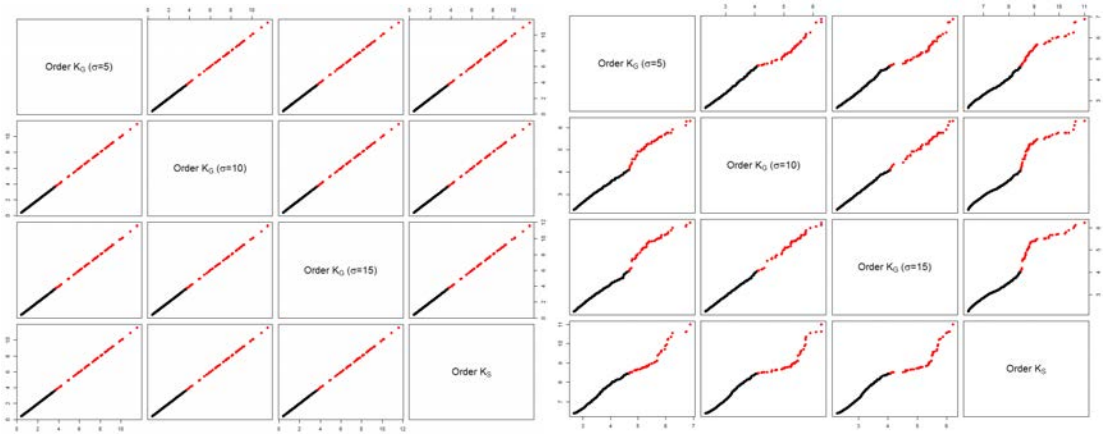


Figure 4.6: Order induced by the entropy estimation for different kernel functions with $\nu = 5\%$ and $n = 2000$. Parametric approach (left) and non–parametric approach (right). The regular curves, corresponding to $X(t)$, in (•) and the detected outliers, corresponding to $Y(t)$, (•).

4.3.4 Shape outlier detection: a single run experiment

The aim of this experiment is to illustrate the performance of the proposed methodology when the atypical data cannot be inferred considering particular extreme points in the curves and under different assumptions about the noise in the observed data. To this aim, a fraction $1-\nu = 90\%$ of $n = 400$ curves comprises the realizations of the following stochastic model:

$$X(t) = sin(t) + cos(t + \varepsilon_l) + a_l + b_l^2, \text{for } l = 1, \ldots, (1-\nu)n, \text{and } t \in [0, 2\pi],$$

where the random coefficients $(\varepsilon_l, a_l, b_l)$ are independent and normally distributed with means $\mu_\varepsilon = 0, \mu_a = 5$ and $\mu_b = 1$, and variances $\sigma_\varepsilon = \sigma_b = 0.25$ and $\sigma_a = 0.2$. The remaining proportion of the data comprises outliers that contaminate the sample accord-

ing to the following stochastic model:

$$Y(t) = sin(t)+cos(t+\varepsilon_l)\frac{1}{2}(sin(2\pi t)+cos(\pi t+\varepsilon_l))+a_l+b_l^2, \text{for } l = 1,\ldots,n\nu, \text{and } t \in [0, 2\pi],$$

where the random coefficients $(\varepsilon_l, a_l, b_l)$ are independent and normally distributed with the same means and variances as in the case of $X(t)$. In Figure 4.7, we show simulated raw data on the left and the corresponding functional data on the right, we use a Gaussian kernel and choose the parameters by cross-validation. In Figure 4.8, we illustrate the outliers captured with the proposed method in red (—), false positives in blue (—) and false negatives in green (—). The parametric approach –Figure 4.8 (left)– captures all the atypical curves in the sample without any false positive, nor false negative finding. The non-parametric approach –Figure 4.8 (right)– shows slightly worse performance incurring four false positive detections and four false negative occurrences. In Table 4.4, we report the TPR, the TNR and the aROC; as can be seen, the proposed methods clearly outperform the other methods in the literature.

Table 4.4: Sensitivity (TPR), specificity (TNR) and the area under the ROC curves (aROC).

Method	TPR	TNR	aROC
MMBD	5.0	89.4	0.452
HMD	12.5	90.3	0.701
RTD	10.0	90.0	0.591
FSD	7.5	89.7	0.645
Entropy-PA	**100.0**	**100.0**	**1.000**
Entropy-NPA	**90.0**	**98.9**	**0.992**

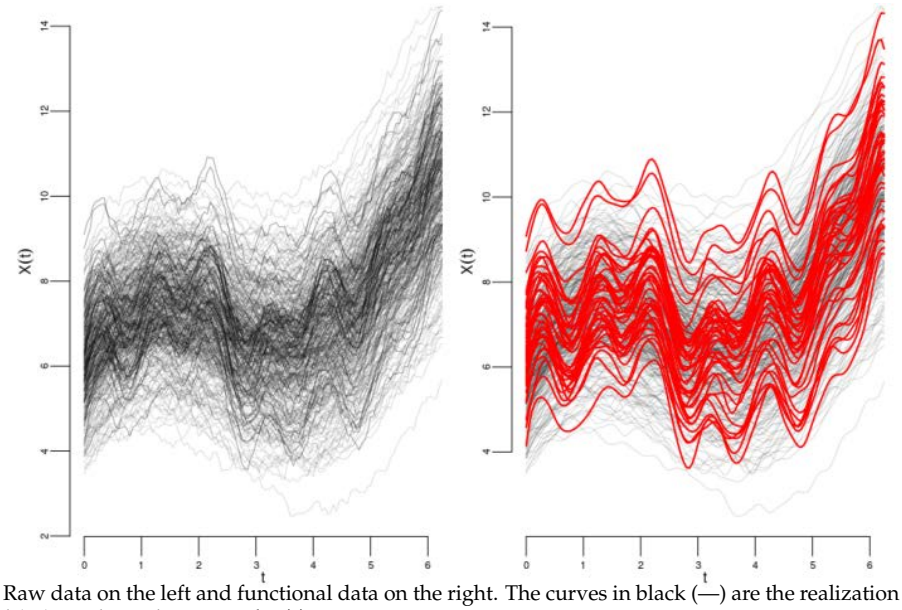


Figure 4.7: Raw data on the left and functional data on the right. The curves in black (—) are the realization of $X(t)$ and paths in red (—) are the realizations of $Y(t)$.

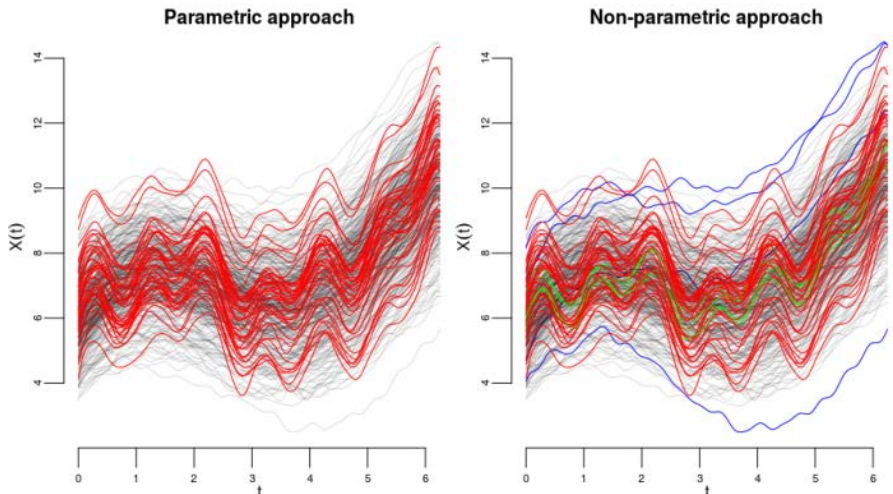


Figure 4.8: Experimental data: in black (—), normal paths corresponding to the realizations of $X(t)$, in red (—), true outlier detected corresponding to the realizations of $Y(t)$ in blue (—) and false negative in green (—).

4.4 Chapter summary

In this chapter we propose a definition of Entropy for stochastic processes, considering a Reproducing Kernel Hilbert Space model to estimate the Entropy from a random sample of realizations of a stochastic process, namely functional data, and introduce two approaches to estimate minimum entropy sets for functional anomaly detection.

We also show the convergence of the parametric Entropy estimation method to the true values through a montecarlo simulation. Morover the order invariance property is studied for both the parametric and non–parametric approach.

In the experimental section, the Monte Carlo simulation illustrates the adequacy of the proposed method in the context of magnitude and shape outliers, outperforming other state of the art methods for functional anomaly detection. In the study of French mortality rates, the parametric and non-parametric approaches for minimum entropy sets estimation show its adequacy to capture anomalous curves, principally associated to the First and Second World Wars and the Influenza episode in 1919. Even though the Gaussian assumptions are not satisfied in this example the parametric approach (PA) behave well in comparison with non parametric approach (NPA).

Chapter 5

An RKHS Autorregressive Hilbertian Model: FA–RKHS

Functional Data Analysis (FDA) deals with objects that can be expressed in the form of functions. In gneneral functional data can be defined as a set of random sample of independent real–valued elements on a compact interval $T = [a, b]$. This random sample is constituted by realizations of a stochastic process $X(t) \in L^2$, where $\mathbb{E}\int_T X^2(t)dt < \infty$, –see Chapter 2 for further details–.

Within the FDA field we can find data structed as independent realizations of a stochastic process, such as in 1.1 (left panel) of Chapter 2; curves that reflect information of a spatial distribution of generating process, such as in 1.1 (right panel) of Chapter 2. Other specific type of functional data is functional time series (FTS). An example regarding fertility rates energy loadings curves is illustrated in Figure 5.1

The structure beyond a FTS set is embedded in the ussual approach. Lets consider $(\Omega, \mathcal{F}, P)$ as the probability space where the random functions of interest are defined, where $\mathcal{F}$ is the σ-algebra in Ω and P a σ-finite measure. We consider random elements (functions) $X(\omega, t) : \Omega \times T \to \mathbb{R}$ in a metric space (T, τ). As usual in the case of functional data, the realizations of the random elements $X(\omega, \cdot)$ are assumed in $C(T)$, the space of real continuous functions in a compact domain $T \subset \mathbb{R}^d$ endowed with the uniform metric. From a practical perspective, one cannot actually observe a functional data set in its entirety. Thus, analysis might be conducted departing from some discrete version of the curve, say $x(t_1), \ldots, x(t_m)$. Each of the real values $x(t_i)$ are measured in an almost contiuous values of the domain $t \in T$, what is called in the literature as raw functional data.

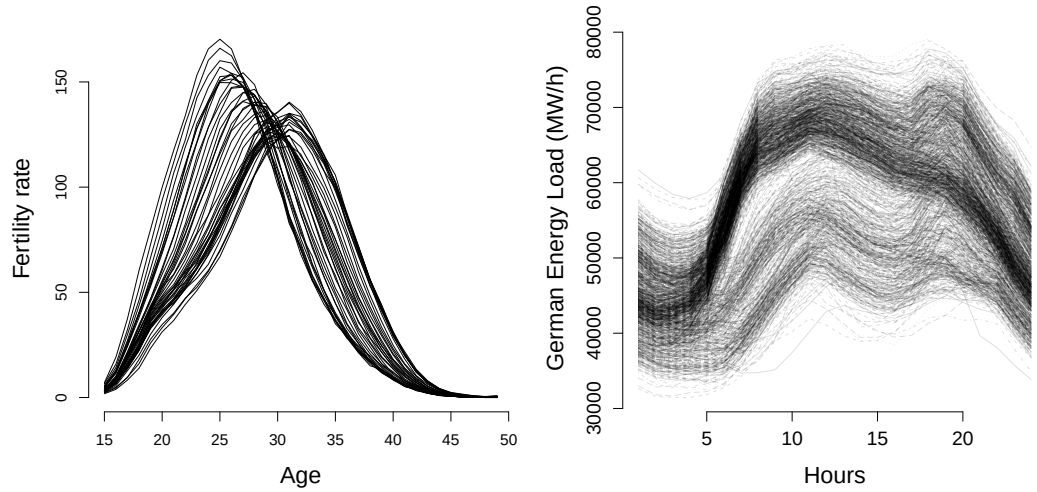


Figure 5.1: Left: Australian fertility rates (1963–2006). Right: German hourly energy loads (01/01/2015–30/05/2018).

When dealing with FTS we must add extra specifications to the previous framework. First, the domain of the random functions in this case is univariate, that is $T \subset \mathbb{R}$. Second, the domain refers to time. This means that the domain were the functions are observed is fixed or random grid of discretized –and in general equally spaced– time points. The frequency of the observed functions depends on how dense is the domain T. One one hand we can observe high–frequency functional time series data such as financial data, i.e.: intra–day prices of a given stock asset, and on the other hand we can observe low–frequency functional time series data, i.e.: the monthly sea surface temperatures.

Third, usually these functions are obtained from a univariate continuous–time stochastic process $Z = Z(t), t \in \mathbb{R}$, and are converted into an stochastic process $X = \{x_i(t), i \in N, t \in [0, \delta]\}$, which is now a discrete-time process that takes values on some functional space. The construction is obtained considering, for a given trajectory of Z observed over the interval $[0, T], T > 0$, the n subintervals of form $I_i = [(i-1)\delta, i\delta], i = 1, \ldots, n$, such that $\delta = T/n$, see Nagbe et al. (2018) for further details.

Fourth, and the most important feature, is the temporal dynamic between the observations. In the case of the well known Berkeley growth data or the Canadian weather data, is assumed that the observations are independent realizations of the same stochastic process. In the FTS context, one can consider that the observations or curves follows a natural order in time. Under this framework of temporal dependence between observations, the assumption of independence is a strong one. In the case of the hourly

energy loadings, illustrated in Figure 5.1 (right panel) is possible to observe that the values of the consumption of energy at the end of the day are highly correlated with the levels of consumption at the begining of the day.

5.1 The autoregressive Hilbertian model: ARH

In the FTS context, when the temporal dynamic of the series is given such that each observation (function) depends stochastically on the previous ones, we say that the stochastic process that generated the FTS is a functional autoregressive process (FAR). One possible approach to model this autoregressive temporal dependence between the curves is to consider the functions, indexed in time, as elements in some separable Hilbert space $\mathcal{H}$ and construct a dynamic system that relates each function with the previous ones in an autoregressive fashion. This dynamic system is called the Autoregressive Hilbertian Model, hereafter ARH. In particular when the temporal dependence if of order 1, which means that the autoregressive equation of the system has one lag, we denominate this ARH as an Autoregressive Hilbertian Model of order 1, or ARH(1). The theoretical framework of the ARH was initially developed by Bosq (2012) and continued by Horváth and Kokoszka (2012) among others.

Following the previous notation, the sequence $\{X_n, -\infty < n < \infty\}$ of zero mean elements of L^2 follows a functional autoregressive proces of order 1, an FAR(1) if,

$$X_n = \Psi(X_{n-1}) + \epsilon_n, \tag{5.1}$$

where $\Psi \in L$, and ϵ_n is a sequence of iid elements in L^2 such that $\mathbb{E}\|\epsilon_n\| < \infty$. For a given ARH(1) such that, $EX_0 = \mu$, $E\|X_0\|^2 < \infty$, the random functions are maps taking values on some (real) separable Hilbert space $\mathcal{H}$. The FAR(1) is as follows,

$$X_n = \mu + \Psi(X_{n-1} - \mu) + \epsilon_n, \tag{5.2}$$

with Ψ a linear bounded (continuous) operator on $\mathcal{H} \mapsto \mathcal{H}$ and $\epsilon = (\epsilon_n, n \in \mathbb{Z})$ a simultaneous withe noise.

5.1.1 Existence

The existence of the ARH is related to i) the compactness of the operator Ψ, and ii) a stationarity condition asociated to Ψ. The operator Ψ is said to be compact if exist a sequence of orthonormal bases, $\{e_i\}$ and $\{g_i\}$ and a convergent sequence $\{\lambda_i\} \in \mathbb{R}$, such that Ψ can be writen as,

$$\Psi(X_n)(t) = \sum_{i=1}^{\infty} \lambda_i \langle X_n, e_i \rangle g_i,$$

for $X_n \in L^2$. If $\sum_{i=1}^{\infty} \lambda_i < \infty$, the operator Ψ is a Hilbert-Schmidt operator. The stationarity condition follows the same logic as for scalar autoregressives models of order 1, AR(1): $y_n = \psi y_{n-1} + \varepsilon_n$. The idea is to impose some restricctions such that the process admits the following expansion

$$y_n = \sum_{k=1}^{\infty} c_k \varepsilon_{n-k},$$

an infinite sum of random elements with zero mean and finite variance. The condition of the AR(1) is $|\psi| < 1$. In the functional case, the condition for operator Ψ is stated in the following Theorem:

Theorem 5.1. *If exists an integer i_o such that $\|\Psi^{i_0}\| < 1$, then there is a unique strictly stationary causal solution to model in Eq. 5.1, and is given by:*

$$X_n = \sum_{k=1}^{\infty} \Psi^i(\varepsilon_{n-i}).$$

The series X_n converges a.s. and in the L^2 norm.

For a formal proof of Theorem 5.1 and further details see Horváth and Kokoszka (2012).

5.1.2 Estimation

Recalling the scalar autoregressive process of order 1, AR(1): $y_n = \psi y_{n-1} + \varepsilon_n$, under the stationarity condition $|\psi| < 1$, the autocorrelation coefficient is $\psi = \gamma_1 \gamma_0^{-1}$, where $\gamma_k = \mathbb{E}[y_n y_{n+k}] = COV(y_n, y_{n+k})$. The sample estimators are obtained by replacing γ_k for the sample autocovariance. The analogous extension for the ARH(1), under Theorem 5.1 is:

$$\Psi = \Gamma_0 \Gamma_1^{-1}, \tag{5.3}$$

where Γ_k is the autocovariance opeartor defined as $\Gamma_k(x) = \mathbb{E}[\langle X_n, x \rangle X_{n+k}]$, and in particular Γ_1 is the lag–1 autocovariance operator. The autocovariance operator Γ admits the following decomposition:

$$\Gamma(x) = \sum_{i=1}^{\infty} \lambda_i \langle x, \nu_i \rangle \nu_i \Rightarrow \Gamma^{-1}(x) = \sum_{i=1}^{\infty} \lambda_i^{-1} \langle x, \nu_i \rangle \nu_i, \tag{5.4}$$

where ν_i is the i^{th} eigenfunction of the operator Γ and λ_j its correspondent associated eigenvalue. Given the expression in (5.4), the inverse of the autocovariance operator is defined if the all the elements of the sequence $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq \lambda_{p+1} = 0$ are positive, but as $\lambda_{p+1} = 0$ the sequence $\{\lambda_i^{-1}\}_{i=1}^{p+1} \to \infty$, does not converge, hence the operator Γ^{-1} is unbounded. Therefore, the identity in Eq. 5.2 is not completely correct given that Γ_1^{-1} is not bounded on the whole $\mathcal{H}$. A empirical solution proposed in Horváth and Kokoszka (2012) is to consider the p most important functional principal components of the operator Γ,

$$\widehat{\Gamma}(x) = \sum_{i=1}^{p} \hat{\lambda}_i \langle x, \hat{\nu}_i \rangle \hat{\nu}_i,$$

where the opoerator $\widehat{\Gamma}^{-1}(x)$ is bounded on the whole space L^2 if $\hat{\lambda}_i$, for $i \leq p$. Consequently the sample autocovariance operator is

$$\widehat{\Gamma}(x) = \frac{N}{N-1} \sum_{k=1}^{N-1} \langle X_k, x \rangle X_{k+1}. \tag{5.5}$$

The sample version of the autocorrelation operator Ψ in Eq. 5.3 is:

$$\widehat{\Psi}_p(x) = \frac{1}{N-1} \sum_{k=1}^{N-1} \sum_{j=1}^{p} \sum_{i=1}^{p} \langle x, \hat{\nu}_j \rangle \langle X_k, \hat{\nu}_j \rangle \langle X_{k+1}, \hat{\nu}_i \rangle \hat{\nu}_i. \tag{5.6}$$

In Theorem 8.7 of Bosq (2012) the author discuss suitable conditions for the relatively fast convergence of the estimated autocorrelation operator, $\|\hat{\Psi}_p - \Psi\| \to 0$.

5.1.3 Prediction

Having observed $X_1, \ldots, X_n$, the question that arises is how to predict X_{n+1}. The dinamics of the ARH(1) model show that the prediction of X_{n+1} is given by the following expression:

$$\hat{X}_{n+1} = \hat{\Psi}(X_n - \mu). \tag{5.7}$$

Firstly, as it is usual in Functional Data Analysis, to represent each function (time series) –which are infinite-dimensional objects by nature–, we need to choose an orthonormal bases of functions $B = \{\phi_1, \ldots, \phi_D\}$, where each ϕ_i belong to some functional subspace $\mathcal{H} \subset L^2$, and then represent each curve by means of a linear combination in $\mathrm{Span}(B)$. Given a discrete curve $\{x(t_i)\}_{i=1}^{m}$, the functional data estimator can be expressed as the sum of basis coefficients times basis functions $\phi_i(t)$ of the form,

$$\tilde{x}(t) = \sum_{i=1}^{m} \alpha_i \phi_i(t).$$

In Ramsay (2006) the authors suggest the use of Fourier basis or the Spline basis system. From another perspective, Antoniadis et al. (2006) develop a methodology based on the functional representation of the FTS in the coodrinates of a Wavelets basis syistem. Other alternative prediction approaches are: i) the optimal expansion of $\Psi(X_n)$ method that is called Predictive Factors, Kargin and Onatski (2008); and ii) the estimation of univariate ARIMA models applied to the scores of the functional principal components of the FTS, Hyndman and Shang (2009). To go deeper in the empirical properties of the forecast with the functional autoregressive model see Didericksen et al. (2012).

Jointly with the *naive* predictor $\hat{X}_{n+1}(t) = X_n$ and the mean predictor or *persistence* $\hat{X}_{n+1}(t) = \mu$, some of these previous methodologies are considered as testing methods of our proposal that is based on the use of a Reproducing Kernel Hilbert Space of basis functions to represent the functional time series. In the next Section we present and deeply discuss this proposal in detail and its advantages.

5.2 An RKHS model for Functional Time Series: FA–RKHS

The Functional Autoregressive of order 1 under a Reproducing Kernel Hilbert Space model, hereafter FA-RKHS is based on the RKHS framework for functional time series. Recalling Section 2.1 of Chapter 2, when we want to represent functional data we face the issue of chosing the *basis expansions*,

$$\tilde{x}(t) = \sum_{i=1}^{m} \alpha_i \phi_i(t),$$

where the set of orthonormal bases of functions, ϕ_i, belong to some functional subspace $\mathcal{H} \subset L^2$. Our choice is to select $\mathcal{H}$ as an RKHS, such that each $\phi_i(t)$ is i^{th} the eigenfunction associated to the positive–definite and symmetric kernel function $K : T \times T \to \mathbb{R}$ that span $\mathcal{H}$. Therefore the functional data estimator can be expressed as

$$\tilde{x}(t) = \sum_{i=1}^{m} \alpha_i K(t, t_i) = \boldsymbol{\alpha}^T \mathbf{k}_t, \tag{5.8}$$

where $\mathbf{k}_t = (K(t_1, t), \ldots, K(t_m, t))$ is the vector of kernel evaluations and the linear combination coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m$ are obtained as the solution of the

following linear system

$$(\gamma \mathbf{I}_m + \mathbf{K})\boldsymbol{\alpha} = \mathbf{X},$$

for $\mathbf{X} = (x(t_1), \ldots, x(t_m))^T$, $\mathbf{I}_m$ an $m \times m$ identity matrix, and $\mathbf{K}$ the Gram matrix with the kernel function evaluations, $[\mathbf{K}]_{k,l} = K(t_k, t_l)$, for $k, l = 1, \ldots, m$. See Chapter 2 for further details.

5.2.1 Estimation of the FA-RKHS

The FA–RKHS model is stated in Eq. 5.9, where Ψ_K is the autocorrelation operator that depends on the kernel function selected to represent the functional time series.

$$\hat{X}_n = \hat{\Psi}_K(X_{n-1} - \mu). \tag{5.9}$$

To estimate the operator Ψ_K we need to define the Covariance and Cross–Covariance operators. Set $T = [0, 1]$ and $\Gamma_0^{(K)} = \mathbb{E}[(X_0 - \mu) \otimes (X_0 - \mu)]$ and $\Gamma_1^{(K)} = \mathbb{E}[(X_0 - \mu) \otimes (X_1 - \mu)]$, then:

$$\Gamma_0^{(K)} \hat{\Psi}_K = \Gamma_1^{(K)}. \tag{5.10}$$

The next step is to estimate μ, $\Gamma_0^{(K)}$ and $\Gamma_1^{(K)}$. In order to do that we use the RKHS functional estimator stated in Eq. 5.8. Given a set of $j = 1, \ldots, n$ functions, the mean function is:

$$\bar{\tilde{x}}(t) = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{m} \alpha_{ji} K(t, t_i) = \sum_{i=1}^{m} \bar{\alpha}_i K(t, t_i), \tag{5.11}$$

where, $\bar{\alpha}_i = \dfrac{1}{n} \displaystyle\sum_{j=1}^{n} \alpha_{ji}$. The autocovariance opertaror $\Gamma_0^{(K)}$ is estimated as follows,

$$\begin{aligned} COV(\tilde{x}(t), \tilde{x}(s)) &= \frac{1}{n} \sum_{j=1}^{n} \left(\sum_{i=1}^{m} (\alpha_{ji} - \bar{\alpha}_i) K(t, t_i) \times \sum_{k=1}^{m} (\alpha_{jk} - \bar{\alpha}_k) K(s, s_k) \right) \\ &= \frac{1}{n} \sum_{j=1}^{n} \left(\sum_{i,k=1}^{m} (\alpha_{ji} - \bar{\alpha}_i)(\alpha_{jk} - \bar{\alpha}_k) K(t, t_i) K(s, s_k) \right), \end{aligned}$$

using the kernel trick, $K(x, y) = \langle K(x, \cdot), K(\cdot, y) \rangle$ the previous the covairance opetator can be written as:

$$COV(\tilde{x}(t), \tilde{x}(s)) = \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i,k=1}^{m}(\alpha_{ji} - \bar{\alpha}_i)(\alpha_{jk} - \bar{\alpha}_k)K(t_i, s_k)\right), \forall\, t, s. \tag{5.12}$$

Analogously the Cross-covariance operator $\Gamma_1^{(K)}$ is estimated by the cross-covariance function,

$$COV(\tilde{x}_l(t), \tilde{x}_g(s)) = \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i,k=1}^{m}(\alpha_{ji} - \bar{\alpha}_i)(\beta_{jk} - \bar{\beta}_k)K(t_i, s_k)\right), \forall\, t, s, \tag{5.13}$$

where $beta_i$ are RKHS expansion coefficients of the functional data estimator $\tilde{x}_g(s) = \sum_{i=1}^{m}\beta_i K(t, t_i)$, for $l, g = 1, \ldots, n$, and $l \neq g$.

5.2.2 Numerical experimients: assesing the predictive performance

In this first numerical experiment we apply our proposed model to obtain point functional predictions in several simulated and real data sets. To test our method we consider the following alternatives approaches:

- ARH–Splines: $\hat{X}_{n+1} = \hat{\mu} + \hat{\Psi}_S(X_n - \hat{\mu})$, Ramsay (2006).
- ARH–Wavlets: $\hat{X}_{n+1} = \hat{\mu} + \hat{\Psi}_W(X_n - \hat{\mu})$, Antoniadis et al. (2006).
- AR–FPCA: $\hat{X}_{n+1} = \hat{\mu} + \sum_{i=1}^{P}\hat{\beta}_{n+1,i}\phi_i$,

  where $\{\phi_i, \ldots \phi_P\}$ are the first P functional principal components of the functional time series, and $\hat{\beta}_{n+1,i}$ is the forecast of the i^{th} FPC score. The forecast is obtainted with a univaraite autoregressive model applied to the socore of each component, –see Hyndman and Shang (2009) for further details–.
- Persistence: $\hat{X}_{n+1} = \hat{\mu}$
- Naive: $\hat{X}_{n+1} = X_n$

For the FA–RKHS model we consider a Gaussian kernel function, $k(t, t_i) = \exp^{\sigma(t-t_i)^2}$, where the parameter σ is defined by grid search, –for further details see Apendix C–. For the rest of the testing methods, the number of basis functions, functional principal components and the smoothing parameter –in the case of the ARH–Splines– are defined through cross validation.

Simulated data sets

To cover all the scenarios that reflect the intrinsic characteristic of the mentioned models above, in the simulated experiment we construct a data set obtained from a FAR(1) process, a data set obtained by a Scalar AR structure applied to the coefficients of some suitable basis functions and a Wiener process. In each case we simulate a functional data set with a sample size of N observations over the domain T, where only the first $N - h$ are input observations to construct each model. Then a prediction of h–steps (in–sample prediction) is constructed. For each simulated experiment we conducted a Monte–Carlo study of $n = 100$ replicates, and we report the total average of the RMSE of the predictions, and the standard error, where the RMSE is computed as:

$$RMSE_h = \sqrt{\int_T (X_{n+h}(t) - \widehat{X}_{n+h}(t))^2 dt},$$

$$Av.RMSE_{h,n} = \frac{1}{n}\sum_{i=1}^{n} RMSE_{h,i}$$

FAR(1) process. We simulate a FAR(1) process following the scheme presented in Eq. 5.1, already implemented in the R–Package `far`. The sample size of each replicate is $N = 100$, sampled at 64 equally spaced points $t = [0, 1]$. In Figure 5.2 we present the last instance of the Monte–Carlo study (left panel), the grid search for the kernel parameter σ (middle panel) and the box–plot showing the $RMSE$ of each of the predicciton methods (right panel).

AR–coefficients. This simulation scheme involves the simulation scalar AR values as the coefficients of some orthonormal basis of functions. The procedure is as follows:

i) Simulate d scalar AR(1) of sample size N, with ϕ_1. $z_i = \mu_d + \phi_{1,d} z_{i-1,d} + \epsilon_i$, with $\epsilon \sim \mathcal{N}(0, 5)$, for $i = 1, \ldots, N$.

ii) Each $z_{i,d} \in \mathbb{R}^d$ and constitutes the coefficients of an orthonormal basis generated by a Legendre polynomial of order d.

iii) Each functional observation $x(t) = \sum_{i=1}^{N}\sum_{k=1}^{d} z_{i,k} P_d(t) + f(t)$, where t is the domain sampled at m equally spaced points and $P_d(t)$ is the Legendre polynomial of degree d, expressed as: $P_d(t) = \dfrac{1}{2^d(d)!}\dfrac{\delta^d}{\delta t^d}[(t^2 - 1)^d]$.

For this experiment we consider, $\phi_1 \sim \mathcal{U}[0.6, 0.7]$, $\mu_d = \{2, 4, 5, 5\}$ and $f(t) = cos(8\pi t) + \varepsilon(t) + c$, where $c \sim \mathcal{N}(0, 0.1)$ and $\varepsilon(t)$ is a withe noise process.

Wiener process. We simulate a Wienner process $X_n(t)$ where $X_n(t) - X_n(s) \approx \mathcal{N}(0, t-s)$, by the central limit theorem. $\mathbb{E}[X_n(t)] = 0$, $V[X_n(t)] = t$, and $COV(X_n(t), X_n(s)) = min(t, s)$.

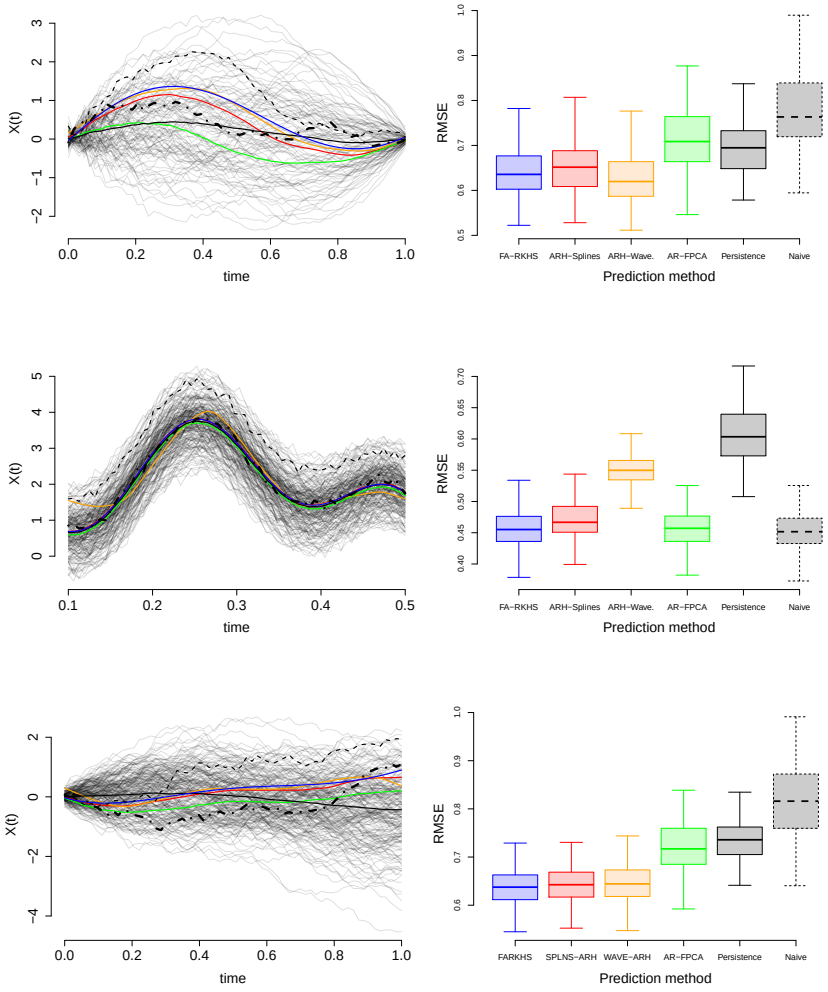


Figure 5.2: By columns: Last instance simulation for each process, the observed function (·····) and the forecasted functions: FA–RKHS (—), ARH–Splines (—), ARH–Wavelets (—), AR–FPCA (—), Persistence (—) and Naive (- - -) (left panels). Comparative boxplot of the predictive methods (right panels). By rows: FAR(1) (upper panel), AR–coefficients (middle panel), Wiener processes (bottom panel).

Table 5.1: Monte–Carlo study: Average RMSE for the h–step ahead forecast for different models –in columns–. Standard errors are reported in parenthesis.

Simulated process	FA–RKHS	ARH–Splines	ARH–Wavelets	AR–FPCA	Persistence	Naive
FAR(1)	0.64	0.65	0.63	0.71	0.70	0.79
	(0.06)	(0.06)	(0.06)	(0.07)	(0.06)	(0.11)
AR–coefficients	0.46	0.47	0.55	0.46	0.60	0.45
	(0.03)	(0.03)	(0.03)	(0.04)	(0.05)	(0.03)
Wiener	0.64	0.64	0.65	0.72	0.74	0.82
	(0.04)	(0.04)	(0.04)	(0.05)	(0.04)	(0.08)

As it can be appreciated in Figure 5.2 and Table 5.1, all the methods present a similar performance, except for the *Persistence* and *Naive*. In particular the proposed FA–RKHS present the lowest average RMSE and standard error. Nevertheless, we cannot conclude that the results are statistically different. The FA–RKHS and the ARH–Splines present similar results in the three Monte–Carlo experiments. While the ARH–Wavelets present better results for the FAR(1) process data set, but a low perfomance in the case of the AR–coefficients. In this data set besides the FA–RKHS the other method that presents good results is the AR–FPCA. This is due to the fact that the simulation scheme reflects the instrinsic characteristics of the model behind the AR–FPCA proposed by Hyndman and Shang (2009). In this sense, what is interesting to mention is that the FA–RKHS performs well under the three simulation schemes. The optimal values sigma for the FA–RKHS model are: $\sigma = \{0.6842; 50; 0.3157\}$ respectively, –see Apendix C–.

Real data examples

For the real data examples we consider two functional data sets. The first one is the Sea Surface Temperature (SST) data set. Each of the curves represent the average monthly sea surface temperature from 01/1950 to 12/2017, measured in the "Niño region": 0–10 degree South. and 90–80 degree West. In the German energy loads (GEL) –illustrated in Figure 5.1– each curve represent the hourly total energy load from 01/01/2015 to 30/05/2018. In both real data examples the objetive is to measure the predictive power of each method, measured in terms of the RMSE. For the SST we consider $h = 35$ (35 years) and for the GEL we set $h = 180$ (six months).

Table 5.2: Average RMSE for each h–step ahead forecast for different models –in columns–. Standard errors are reported in parenthesis.

Simulated process	FA–RKHS	ARH–Splines	ARH–Wavelets	AR–FPCA	Persistence	Naive
SST (h=35)	0.94	0.98	1.00	1.00	1.40	0.98
	(0.66)	(0.71)	(0.67)	(0.70)	(0.99)	(0.70)
GEL (h=180)	3163.35	3419.04	4567.52	6284.73	5357.33	7706.93
	(2295.66)	(2490.42)	(2209.08)	(3610.24)	(4518.61)	(2874.23)

Up to now, one of the constributions of this Chapter is the proposal of a new family of set of basis functions to estimate an autoregressive Hilbertian model for functional time series. In this sense, the results presented above –see Figures 5.2, 5.3 and Tables 5.1, 5.2– show similar performance among the different models considered. Nevertheless, the construccion of the FA–RKHS model, proposed in this Chapter, entails a stability property of the basis expansion coefficients that allows us to construct perdictive confindence bands for the forecasted functions $\widehat{X}_{n+h}$. This property and the methodology for such bands are detailed in the next Section.

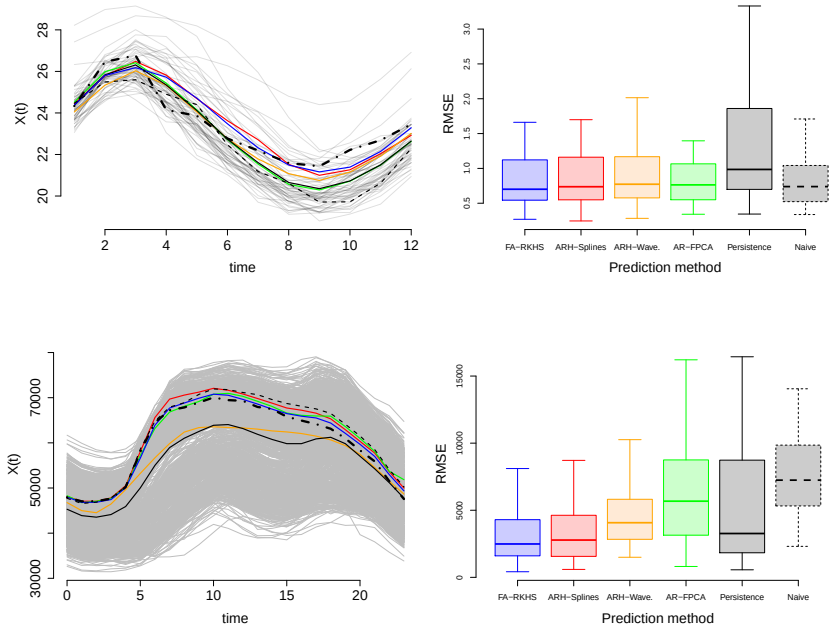


Figure 5.3: By columns: Functional data set, the observed function (·····) and the forecasted functions: FA–RKHS (—), ARH–Splines (—), ARH–Wavelets (—), AR–FPCA (—), Persistence (—) and Naive (- - -) (left panels). Comparative boxplot of the predictive methods (right panels). By rows: Sea Surface Temperature (upper panel) and German Energy Loadings (bottom panel).

5.3 Confidence bands

Prediction is one of the main objectives of time series analysis. In terms of uncertainty, prediction intervals bring more information about the future values of a random variable than point forecasts. Prediction intervals give a set of values that the realization of the future random variable could take, conditional to past information and given a certain probability. In the functional time series context, given the functional nature of the observations, the predictive confidence intervals take the form of predictive confidence bands.

5.3.1 Constructing the confidence bands

Pointwise vs simultaneous inference

There are at least two alternatives to conduct inference in the functional context: i) pointwisely or ii) simultaneously. In functional data is important to asses the prediction over the entire domain T of the functions rather than doing it for a single point $t_0 \in T$. To motivate the difference between pointwise and simultaneous inference, lets consider the functional prediction of a particular functional model, $\widehat{X}_{n+1}$. The objective is to

define the functional statistics, $L^s_{n+1}(t)$ and $U^s_{n+1}(t)$ such that the band constituted by $\{[L^s_{n+1}(t), U^s_{n+1}(t)] : t \in T\}$ fully contains the prediction $\widehat{X}_{n+1}$ with a probability of $1-\nu$:

$$P(\widehat{X}_{n+1} \in [L^s_{n+1}(t), U^s_{n+1}(t)], \forall t \in T) = 1 - \nu$$

A methodology that consider pointwise predictive intervals $\{[L^p_{n+1}(t), U^p_{n+1}(t)]\}$ with $1-\nu$ level for each $t \in T$ and then joint each interval in a confidence band will not satisfy that $P(\widehat{X}_{n+1} \in [L^p_{n+1}(t), U^p_{n+1}(t)], \forall t \in T) = 1 - \alpha$[1]. Even though the pointwise predictive bands are a valid inferencial method, in general their coverage is less thant $1-\nu$, which can mislead the conclusion in terms of confidence of the prediction. See Degras (2017) and Wolf and Wunderli (2015) for a deeper discussion.

Parametric vs. non–parametric inference

One important issue when constructing prediction intervals or regions, is the assumptions made with respect to the distribution of the innovations of the process. The standard approach is to assume Gaussian innovations, which generate prediction intervals centered on the conditional expectation function, and do not consider the uncertainty derived from the parameter estimation. In general Gaussian assumptions do not constitute a good probabilistic framework when dealing with real time series data, such as financial data. Hence some alternative approach should be used to address this issue.

Usually the analytical derivation of the distribution of an estimator is very difficult or even impossible and only a sampling approximation is available. In that context the bootstrap technique arose, as a method that allows to get an approximation of the estimator distribution throughout drawing with replacement random samples from the empirical distribution function. The bootstrap technique presents several advantages: i) it is no necessary to make any assumption about the distribution of the population from which the sample was obtained; ii) is an easy technique to implement beyond the complexity of the statistic of interest and iii) the statistics obtained are consistent under conditions shown latter on, see Efron and Tibshirani (1994).

The seminal framework under the development of bootstrap techniques is the *iid* case. When we analyze data that present some structural dependence, such as time series data, the iid bootstrap techniques lead to inconsistent statistics. There are several methodologies oriented to tackle this problem and proposed bootstrap techniques for time series. For an extensive review of the topic see Kreiss and Lahiri (2012); Kreiss

[1]Supraindexes p and s refer to pointwise and simultaneous respectively

and Paparoditis (2011); Härdle et al. (2003). In the functional time series context several contributions have been made to the field, see e.g. Shang (2018); Paparoditis et al. (2018).

In this Chapter we consider an extension of the bootstrap methodology proposed by Pascual et al. (2004); Fresoli et al. (2014) in the univariate and multviarate framework respectively, to the functional context. The authors in Pascual et al. (2004) proposed a model based bootstrap methodology, hereafter PRR, which is summarized below:

Given the stationary $AR(p)$ process,

$$y_n = \phi_1 y_{n-1} + \phi_2 y_{n-2} + \cdots + \phi_p y_{n-p} + \varepsilon_n,$$

(i) obtain the residuals of the estimation using $\hat{\varepsilon_n} = y_n - \sum_{i=1}^{p} \hat{\phi}_i y_{n-i}$, with an appropiate rescale procedure, see Stine (1987) and let $\hat{\mathbf{F}}_\varepsilon$ be the its empirical cummulative distribution function;

(ii) resample from $\hat{\mathbf{F}}_\varepsilon$ and construct the bootstrap series of length N using $y_n^* = \hat{\phi}_0 + \hat{\phi}_1 y_{n-1}^* + \hat{\phi}_2 y_{n-2}^* + \cdots + \hat{\phi}_p y_{n-p}^* + \hat{\varepsilon}_n^*, \;\; n = 1, \ldots, N$, given a set of initial values $\mathbf{y}_0^* = \{y_{-p+1}^*, \ldots, y_0^*\}$;

(iii) estimate $\hat{\phi}^*$;

(iv) generate a bootstrap forecast, $y_{n+k}^* = \hat{\phi}_0^* + \sum_{j=1}^{p} \hat{\phi}_j^* y_{n+k-j}^* + \hat{\varepsilon}_{n+k}^*$, fixing the last p observations of the series;

(v) Repeat steps (ii)–(iv) B times to otbain the bootstrap replicates.

This methodology present the advantage that does not require the backward representation of the process and can be adapted to non–linear time series. Moreover the procedure is straightforward to apply and computationally efficient. To see a formal derivation of this procedure and the asymptotic properties of the bootstrapped estimators $\hat{\phi}^*$, see Pascual et al. (2004).

Given the ARH(1) process: $X_n = \mu + \Psi(X_{n-1} - \mu) + \epsilon_n$ the functional extension of the PRR procedure is as follows:

(i) Estimate: $\hat{X}_n = \hat{\mu} + \hat{\Psi}(X_{n-1} - \hat{\mu})$ with a functional model, in this work we consider the FA–RKHS model;

(ii) obtain the residuals: $\hat{\epsilon}_n = \hat{X}_n - X_n$;

(iii) resample from $\hat{\epsilon}_n$, and construct the functional bootstrap series $X_n^* = \hat{\Psi}_K^*(X_{n-1}) + \hat{\epsilon}_n^*$ fixing the first initial curve $X_0^* = X_0$;

(iv) estimate $\hat{\Psi}_K^*$;

(v) obtain the h step ahead forecast $X_{n+h}^* = \hat{\Psi}_K^*(X_{n-1+h}^*)$;

(vi) repeat steps (iii)–(v) to obtain the B bootstrap replicates $X_{n+h}^{*(1)}, \ldots, X_{n+h}^{*(B)}$.

5.3.2 ν–Minimum Entropy Sets

To construct the predictive confidence bands we use the B bootstrap replicates –obtained in step (vi) of the bootstrap procedure– for each forecast horizon h. To this aim we consider the concept of ν–minimum–entropy set (MES) detailed in Section 4.2 of Chapter 4. For a given h and for each bootstrap replicate we obtain its functional estimator, solving the regularitazion problem defined in Section 2.1 of Chapter 2, –see Eq.2.3–. The solution to this problem lets us define the RKHS representation of each bootstrap replicate as:

$$\tilde{X}_{d,h}^*(t) = \sum_{j=1}^{d} \sum_{i=1}^{m} \alpha_i l_j v_{i,j} v_{m+1,j} = \sum_{j=1}^{d} z_j^* e_j, \tag{5.14}$$

where $e_j = \sqrt{l_j} v_{m+1,j}$, $z_j^* = \dfrac{\sqrt{l_j}}{\sqrt{m}} \displaystyle\sum_{i=1}^{m} \alpha_i v_{i,j}$, $d < m+1$ and $\tilde{X}_{d,h}^*(t) \in \mathcal{H}_d \subset \mathcal{H}$, –see Section 2.1 of Chapter 2 for further details–. We identify each functional bootstrap replicate in the sample $\tilde{X}_{d,h}^*(t)$ with a vector $\mathbf{z}_b^* = (z_{1,b}^*, \ldots, z_{d,b}^*) \in \mathbb{R}^d$ for $b = 1, \ldots, B$.

Once we obtain the functional representation of each bootstrap replicate $Z^* = \{z_1^*, \ldots, z_B^*\} \in \mathbb{R}^d$, that admits a continuous density function f_Z, we define $H_\alpha(A_{Z^*}^{(h)}) = \frac{1}{1-\alpha} \log \left(\int_A^{(h)} f_Z^\alpha(\mathbf{z}^*) d\mathbf{z}^* \right)$ to be the entropy of the Borel–set $A^{(h)}$ with respect to the measure F_Z^*, and h indicate the forecast horizon. Then, the ν–minimum–entropy set (MES) is formally defined as:

$$\text{MES}_\nu(Z^*) := \{\arg\min_{A^{(h)} \subset \mathbb{R}^d} H_\alpha(A_{Z^*}^{(h)}) \text{ s.t. } P(A^{(h)}) \geq 1 - \nu\},$$

which is equivalent to a ν-high density set (HDS) Hyndman (1996) formally defined as $\text{HDS}_\nu(Z^*) = \{\mathbf{z}^* \in \mathbb{R}^d | \, f_Z^*(\mathbf{z}^*) > c_\nu\}$, where c_ν is the largest constant such that $P(\text{HDS}_\nu(Z^*)) \geq 1 - \nu$, for $0 < \nu < 1$. We are able then to define the simultaneous ν-Predictive Confidence Band.

Definition 5.1 (ν-Predictive Confidence Band)**.** A set $\mathcal{B}$ is a $1-\nu$ prediction band for X_{n+h} if it verifies

$$\mathbb{P}\{X_{n+h} \in \mathcal{B}_\nu\} \geq 1-\nu$$

The $\mathcal{B}_\nu$ with the νB curves whose RKHS representation belongs to $A^{(h)}$, that is:

$$X_{n+h}^{*(b)} \in \mathcal{B}_\nu \iff \mathbf{z}_b^* \in A^(h),$$

Once we define the set A we recover the infinite–dimesional object and obtain the predictive band in the original representation of the curves.

Estimating the ν–Minimum Entropy Sets

- **Parametric–approach (PA)**. Under this approach we assume that $f_Z^*(\mathbf{z}^*, \boldsymbol{\theta})$ is a suitable probability model for the random sample $\mathbf{z}_1^*, \ldots, \mathbf{z}_B^*$, then we estimate by robust maximum likelihood (RML) the parameters $\boldsymbol{\theta}$. As in Chapter 4, we consider $f_Z^*(\mathbf{z}^*, \boldsymbol{\theta})$ to be the normal density, and then RML estimated parameters are $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$, the robust mean vector and covariance matrix respectively. The estimated set MES_ν is defined trough the following expression

  $$\text{MES}_\nu(S_n) = \{\mathbf{z}^* \in \mathbb{R}^d | (\mathbf{z}^* - \widehat{\boldsymbol{\mu}})^T \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{z}^* - \widehat{\boldsymbol{\mu}}) \leq \chi_d^2(\nu)\},$$

  where $\chi_d^2(\nu)$ is the $1-\nu$ quantile of a Chi–square distribution with d–degrees of freedom. Then if the coefficient $\mathbf{z}_b$, lies inside this ellipsoid, we say that the functional datum belongs to the predictive confidence band $\mathcal{B}_\nu$.

- **Non–parametric–approach (NPA)**. This approach involves a more flexible assumption on the distribution of $f_Z^*(\mathbf{z}^*)$. In particular, we estimate the ν–MES solving the One–Class Neighbor Machine problem defined in Subsection 4.2.2 of Chapter 4. The solution to that problem, ρ^*, leads to the following decision function

  $$D(\mathbf{z}) = \text{sign}(\rho^* - \hat{h}_\alpha(\Delta_\mathbf{z})),$$

  where $\hat{h}_\alpha(\Delta_{\mathbf{z}^*})$ is the δ-Local α-Entropy, and $D(\mathbf{z}^*) = +1$ if $\mathbf{z}^*$ corresponds to the (ν) proportion of curves that belongs to a low entropy (high density) set, see 4 and Muñoz and Moguerza (2006) for further details.

Example 5.1. Simulated data set: AR–coefficients ($B = 1000, h = 1$). $\mathcal{B}_\nu$ and ν–MES. In this example we consider the AR–coefficient simulated data set to show the construction of the predictive confidence bands $\mathcal{B}_\nu$. Once we obtain the B bootstrap replicates of the prediction –step (vi) of the bootstrap procedure–, we project them into $\mathcal{H}_d \subset \mathcal{H}$ and

compute the ν–MES, $A_\nu^{(h)}$, for different values of $\nu = \{0.05, 0.1, 0.2\}$, that is, a confifence of $95\%, 90\%, 80\%$ respectively.

Under the parametric approach this represent the concentric ellipsoids for the different values of ν. Given the values of the ellipsoid in $\mathcal{H}_d$, let say $\mathbf{z}^{(e)}$, we recover the infinite–dimensional object by multiplying $\mathbf{z}^{(e)}$ for the basis functions e, using the identity in Eq- 5.14:

$$\mathcal{B}_\nu = \left\{ \sum_{j=1}^{d} \mathbf{z}_{j,b}^{(e)} e_{j,b} \right\}_{b=1}^{B}$$

In the non–parametric approach we take the convex hull of all the $\mathbf{z}_b^* \in A_\nu^{(h)}$. The predictive confidence band in this case will be constituted by the convex hull of all curves associated to its RKHS representation $\mathbf{z}_b^*$, applying Definition 5.1. As can be appreciated in Figure 5.4 the Bands are centered at the forecasted function. In this example the non–parametric approach leads to predictive confidence bands with less uncertainty than the paramtric approach, but with a lower level of coverage.

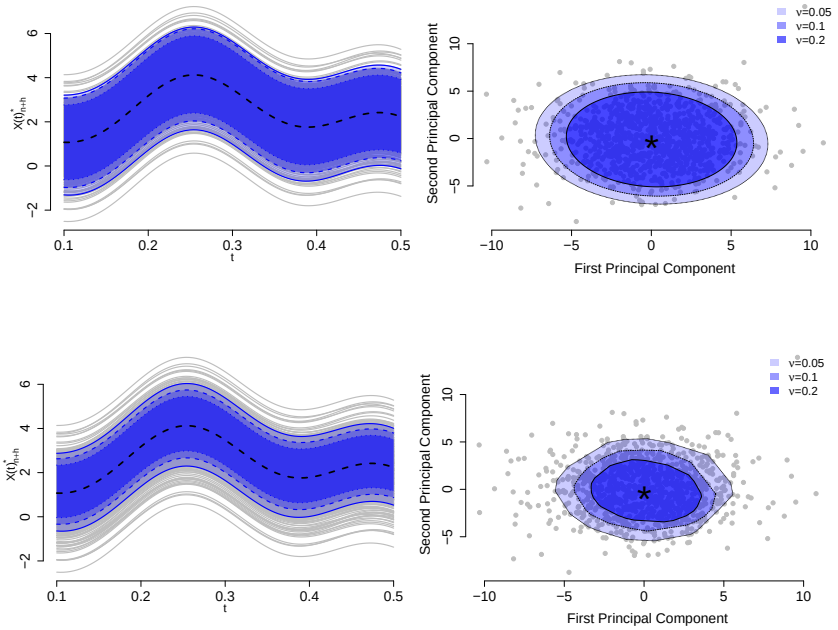


Figure 5.4: Illustration of Example 5.1. The $\mathcal{B}_\nu$ (left) and ν–Minimum Entropy Sets (right) using the Entropy-PA (upper panels) and the Entropy-NPA (bottom panels), for different values of $\nu = \{0.05, 0.1, 0.2\}$. In (- - -) and ($\star$) the forecasted function respectively.

• • •

5.3.3 Theoretical justificaciton

The construccion of the predictive confidence bands is based on the definition of RKHS projection stability detailed in definitions 5.2 and 5.3 and Theorem 5.2, detailed below.

Definition 5.2 (ϵ–perturbed curve)**.** Let $X_n(t)$ be a sample curve, $X_n^{\epsilon}(t)$ is a ϵ–perturbed curve of $X_n(t)$ if:

$$\frac{|X_n(t_i) - X_n(t_i)^{\epsilon}|}{|X_n(t_i)|} < \epsilon, \text{for } i = 1, \ldots, m.$$

Definition 5.3 (RKHS projection stability)**.** Let $\mathbf{z}_n = (z_{1,n}, \ldots, z_{d,n}) \in R^d$ the RKHS representaion of $X_n(t)$, and $\mathbf{z}_n^{\epsilon} \in R^d$ the RKHS representaion of $X_n^{\epsilon}(t)$, then $\mathbf{z}_n$ is ϵ–stable if:

$$\frac{|\mathbf{z}_n^j - \mathbf{z}_n^{j,\epsilon}|}{|\mathbf{z}_n^j|} < \epsilon, \text{for } j = 1, \ldots, d.$$

Theorem 5.2 (RKHS projection stability)**.** *Under the conditions defined in Eq. 2.8, $\mathbf{z}_n$ is ϵ–stable.*

For a formal proof of Theorem 5.2 see Muñoz and González (2010).

The implication of the RKHS projection stability is that two curves (functions) whose RKHS representation is "close" given the metric induced by the Kernel operator in $\mathcal{H}_d$ present a "similar" temporal dynamic in the infinte–dimensional representation space, Muñoz and González (2010). In this sense the continuity of the integral operator, $I_K \int_T K(\cdot, t)X(t)dt$ brings the sufficient regularity properties. Consider the *evaluation mapping* $\delta_t(f) = f(t)$ that assigns a real number to each function. In general, the evaluation functional is not continuous, which implies that $\delta_t(f_n)x \not\to \delta_t(f)$ when $n \to \infty$, even when $f_n \to f$; showing that Hilbert spaces can contain functions that far from being smooth, –see Ramsay (2006)–.

One of the characteristics of reproducing kernel Hilbert spaces is that are the only ones where the evaluation functional $\delta_t(f)$ is continous, which means that the functions in the space are well-behaved. As this condition is not satisfied fot any Hilbert space $\mathcal{H}$, states the advantage of considering our RKHS framework to construct well-behaved and mathematically founded predictive confidence bands.

5.3.4 Numerical experimients: making inference with the predictive bands

For this numerical section we use the same functional data sets consider in Section 5.2.2. The objective of these experiments is to study the coverage of predictive confidence bands proposed under the FA–RKHS model, for different levels of $1 - \nu$ with

$\nu = \{0.05, 0.1, 0.2\}$, namely a nominal coverage of $\{95\%, 90\%, 80\%\}$. For all the cases we use a Gaussian Kernel function where the paremeter σ is defined over a grid search.

To show a testing method of our proposal of constructing predictive confidence bands we present the naive methodology. The naive bands are constructed pointwisely assuming a Gaussian distribution as follows,

$$L_{n+h}^{(naive)}(t) = \hat{X}_{n+h} - Z_{(1+\nu)/2}\sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(X_{n+h}^{(b)}(t_i) - \bar{X}_{n+h}^{(b)}(t_i)\right)^2}, \text{ for } j = 1, \ldots, m$$

$$U_{n+h}^{(naive)}(t) = \hat{X}_{n+h} + Z_{(1+\nu)/2}\sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(X_{n+h}^{(b)}(t_i) - \bar{X}_{n+h}^{(b)}(t_i)\right)^2}, \text{ for } j = 1, \ldots, m$$

The average empirical coverage measures the average number of times that the point forecast $\hat{X}_{n+h}$ is inside the band. The family wise k error (FWKE) is a more flexible measure that allows that in a given percentage k of the whole domain the point forecast (curve) can be out of the band. For this experiment we consider $k = 10\%$. Last, the amplitude is the area contained inside the bands for a given h computed as follows,

$$Amp_h = \int_T (U_{n+h}(t) - L_{n+h}(t))dt$$

In tables 5.3 we report the avareage and standard error –in parenthesis– of the coverage (Empirical), the FWKE coverage, and amplitude (Amp) of the bands for a forecast horizon $h = 1, \ldots, H$. For the simulated data set we consider a sample size $N = 1000$ and $H = 500$. For the Sea Surface Temperature (SST) the sample size is $N = 57$ and $H = 35$ (35 years). For the German Energy Loads (GEL) $N = 1246$ and $H = 180$ (six months).

As it can be appreciated in Table 5.3 the methodology proposed to costruct predictive confidence bands present good performance in both, simulated and real functional time series. In particular, and as it was expected, the level of empirical coverage for the ν–Entropy Bands is (far) higher than the naive coverage. As is also expected the amplitude increase as we increase the level of nominal coverage.

Table 5.3: Empirical coverage, FKWE and amplitud for different nominal coverages $1-\nu$ in columns. In rows, the average metrics for the Entropy paramertic approach (E–PA), Entropy non–parametric approach (E–NPA) and naive approach, for different functional time series (standard-error reported in parenthesis).

		Nominal: 80%			Nominal: 90%			Nominal: 95%		
FTS	Metric	Emprical	FKWE (10%)	Amp.	Emprical	FKWE (10%)	Amp.	Emprical	FKWE (10%)	Amp.
FAR(1)	E–PA	0.85	0.96	1.32	0.88	0.93	1.41	0.94	0.98	1.49
		(0.368)	(0.219)	(0.183)	(0.516)	(0.191)	(0.092)	(0.468)	(0.174)	(0.079)
	E–NPA	0.87	0.94	1.02	0.88	0.97	1.073	0.92	0.97	1.12
		(0.454)	(0.223)	(0.091)	(0.416)	(0.183)	(0.056)	(0.476)	(0.168)	(0.093)
	Naive	0.43	0.52	2.12	0.53	0.65	2.57	0.75	0.79	2.95
		(0.354)	(0.465)	(0.021)	(0.298)	(0.346)	(0.027)	(0.272)	(0.1198)	(0.044)
AR–coef.	E–PA	0.82	0.96	1.13	0.78	0.97	1.14	0.84	0.98	1.14
		(0.386)	(0.196)	(0.083)	(0.416)	(0.171)	(0.082)	(0.368)	(0.14)	(0.087)
	E–NPA	0.77	0.96	1.03	0.78	0.97	1.077	0.82	0.97	1.11
		(0.422)	(0.196)	(0.071)	(0.416)	(0.171)	(0.066)	(0.386)	(0.171)	(0.083)
	Naive	0.46	0.55	2.761	0.73	0.85	2.89	0.85	0.89	3.15
		(0.394)	(0.457)	(0.022)	(0.285)	(0.335)	(0.026)	(0.242)	(0.196)	(0.034)
Wiener	E–PA	0.91	0.95	4.96	0.92	0.96	4.99	0.93	0.97	5.01
		(0.287)	(0.196)	(0.403)	(0.272)	(0.219)	(0.413)	(0.287)	(0.171)	(0.409)
	E–NPA	0.85	0.85	4.15	0.88	0.90	4.42	0.91	0.96	4.62
		(0.358)	(0.326)	(0.852)	(0.326)	(0.301)	(0.310)	(0.287)	(0.196)	(0.349)
	Naive	0.36	0.55	2.01	0.63	0.75	2.57	0.80	0.88	3.07
		(0.282)	(0.451)	(0.05)	(0.485)	(0.435)	(0.066)	(0.402)	(0.326)	(0.077)
SST	E–PA	0.9	0.9	6.74	0.9	0.9	6.74	0.9	0.9	6.74
		(0.307)	(0.307)	(0.736)	(0.307)	(0.307)	(0.561)	(0.307)	(0.307)	(0.568)
	E–NPA	0.80	0.85	4.25	0.90	0.9	4.86	0.85	0.90	5.27
		(4.699)	(7.712)	(0.852)	(5.914)	(9.013)	(1.180)	(5.351)	(7.772)	(1.341)
	Naive	0.4	0.5	2.33	0.65	0.7	3.01	0.7	0.75	3.58
		(0.252)	(0.312)	(0.05)	(0.489)	(0.470)	(0.065)	(0.470)	(0.444)	(0.096)
GEL	E–PA	0.876	0.888	3.45	0.876	0.884	3.47	0.876	0.896	3.44
		(0.330)	(0.315)	(0.550)	(0.330)	(0.320)	(0.565)	(0.330)	(0.305)	(0.523)
	E–NPA	0.818	0.853	1.65	0.893	0.93	2.12	0.94	0.96	2.52
		(0.469)	(0.112)	(0.082)	(0.494)	(0.132)	(0.078)	(0.451)	(0.500)	(0.325)
	Naive	0.276	0.348	1.66	0.392	0.412	2.02	0.436	0.472	2.30
		(0.247)	(0.277)	(0.225)	(0.289)	(0.293)	(0.269)	(0.246)	(0.172)	(0.131)

5.4 Chapter summary

In this chapter we present a new autoregresive Hilbertian model for functional time series. Based on a reproducing kernel Hilbert space framework, the first contribution is to develop a new family of basis functions to estimate the autocorrelation operator Ψ and to predict an entire new function for the whole domain. Throughout several Monte–Carlo studies, we show the performance of the proposed model, in terms of the root mean squared error, against well known prediction methodologies for functional time series.

In a second stage we tackle the issue of constructing predictive conifidence bands for the point forecast. We present a discussion related to the pointwise and simultaneous inference approaches to construct the predictive bands. Our proposed methodology is based on a model–based bootstrap approach for functional time series, which is an extension of the PRR bootstrap procedure. We theoretically justify our proposal based on the continuity of the integral operator, noticing the advantage of the reproducing kernel

Hilbert space framework over other approaches.

We study the pefromance of our simultanous predictive confidence bands procedure, in terms of the emprical coverage, in several simulated and real functional time series data sets.

Chapter 6

Domain selection for functional data

In the era of big data it is becoming more and more common to observe data that arise in the structure of curves almost continuously observed over a grid of discrete time points. Satellite pixel–images evolving during year, or households electricity consumption curves recorded almost continuously during the day are some examples of what is called nowadays functional data. One of the challenges with Functional Data (FD) is the difficult implementation of inferential techniques, since the objects under study are infinite dimensional by nature (see Zhang et al. (2007); Horváth and Kokoszka (2012) and reference therein for a review of statistical analysis with FD). In this context, feature selection techniques has captured the attention of researchers in the field, since allows to reduce the dimension of functional data facilitating the use inferential techniques with curves. Recent works in the literature concerning the use of *domain selection methods* – an appealing way to extract features with FD– with inferential pourposes can be seen in Pini and Vantini (2017); Fraiman et al. (2016); Berrendero et al. (2016) and Hall and Hooker (2016). In the aforementioned works, feature selection can be understood in two possible ways: (i) the selection of isolated points in the domain of the functions; or (ii) the selection of intervals in the domain of the curves, that suffices to implement the inferential methods of interest.

Even though *domain selection* can be applied to many inferential problems with FD, in this chapter we focus in classification problems; where the goal is to learn to infer about the class–membership of a curve. In particular, and motivated by empirical examples, there are reasons to think that the *classification problem can be boosted droping out redundant information* between the classes of functions. Therefore, the problem we tackle in this chapter can be posed in the following way: Let Y be a random latent variable defined in the set $\{1, \ldots, k\}$ –describing the *state of the nature*–; for any classification

function $h \in \mathcal{C} : L^2(T) \to \{1, \ldots, k\}$, where $\mathcal{C}$ is a large set of classification models, we are interested in *learn* the interval $\Theta = (\theta_1, \theta_2) \subset T$, in the following way

$$(\theta_1^*, \theta_2^*) := \arg\max_{\theta_1, \theta_2} P(h_\Theta(X_Y) = y \,|\, Y = y), \tag{6.1}$$

for $y = 1, \ldots, k$, where $X_Y(t) \in L^2(T)$, is a stochastic process that depends upon the state of the nature described by the class latent variable Y and $h_\Theta : L^2(\Theta) \to \{1, \ldots, k\}$ denotes the use of h but only in the selected domain Θ. Therefore the goal of domain selection is to learn a suitable interval Θ that maximizes the probability of correctly classify each curve independently from the classification model h.

To this aim, we introduce the concept of *Kullback-Liebler* divergence curve KL_C and it empirical counterpart. Moreover, we also propose and alternative –and computationally cheaper– approach for domain selection based on a related curve named as *common support curve* S_C. The contributions of this chapter can be summarized as follows:

- We develop a methodological framework for domain selection which is model–free by introducing two relevant concepts, namely KL_C and S_C.
- We propose estimators for both curves and asses the reliability of them by using numerical experiments.
- We also prove, using simulations and real data examples, that *domain selection* effectively improves the performance of a large set of different classification methods, at the same time we also reduce the computational burden–time and memory–for all of them.

The rest of the chapter is organized as follows. In Section 6.1 we introduce the general framework for domain selection in the context of functional time series data. In Section 6.2 we formally introduce the KL_C and S_C curves and its estimation counterparts. Section 6.3 is devoted to the numerical and empirical analysis, and Section 6.4 concludes our work.

6.1 General Framework

6.1.1 Functional time series

For the sequel, we will consider functional data as realizations of a continuous stochastic process $\{X(t, \omega) : t \in T, \omega \in \Omega\}$, suitable defined in a probability space $(\Omega, \mathcal{F}, P)$. The random function $X(t)$ is $\mathcal{F}$-measurable in Ω, and once the process has been realized

the resulting collection of real numbers $x(t)$ is a functional datum. As usual in the case of functional data, we assume $X(t) \in L^2(T)$, the space of square integrable real continuous functions in a compact domain $T \subset \mathbb{R}$.

Since the full realizations of a process $X(t)$ is unobservable for many practical reasons, the analysis of FD might be conducted departing from some discrete version of each curve, say $x(t_1), \ldots, x(t_n)$, and we call these measurements –following the FD literature jargon– as raw functional data. In particular, when the domain T represent time we refer to the collection of raw functional data as Funtional Time Series (FTS) Hörmann and Kokoszka (2012). Therefore, the first step when working with FTS is the representation consisting in recoverer $x(t)$ from its discrete version $x(t_1), \ldots, x(t_n)$. To this aim, and following the usual approach in Functional Data Analysis Ramsay (2006), we must choose an orthonormal basis of functions $B = \{\phi_1, \ldots, \phi_d\}$, where each ϕ_i belong to some functional subspace $\mathcal{H} \subset L^2(T)$, and then represent each curve by means of a linear combination in the $\mathrm{Span}(B)$. Our choice in this chapter is to consider $\mathcal{H}$ as a Reproducing Kernel Hilbert Space (RKHS) of functions Berlinet and Thomas-Agnan (2011). In this case, the elements in the spanning set B are the eigenfunctions associated to the positive-definite and symmetric kernel function $K : T \times T \to \mathbb{R}$ that span $\mathcal{H}$. It can be shown –see Chapter 2– that after solving a *regularization problem*, each functional datum can be approximated by

$$\tilde{x}(t) = \sum_{i=1}^{n} \alpha_i K(t, t_i), \tag{6.2}$$

where $K(t, t_i)$ is the kernel evaluation and the coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ are obtained solving the linear system:

$$(\gamma \mathbf{I}_n + \mathbf{K})\boldsymbol{\alpha} = \mathbf{x}, \tag{6.3}$$

where γ is a regularization parameter (usually fitted by cross–validation), $\mathbf{x} = (x(t_1), \ldots, x(t_n))^T$, $\mathbf{I}_n$ is an $n \times n$ identity matrix, and $\mathbf{K}$ the Gram matrix with the kernel function evaluations, $[\mathbf{K}]_{i,j} = K(t_i, t_j)$, for $i, j = 1, \ldots, n$.

6.1.2 Domain selection to remove reduntant information

The high–dimensionality and complexity of FD poses important challenges in supervised classification problems involving FTS. In such contexts, *domain selection* plays an important role in the sense that is a method developed to extract the features that reflects the *differences* between the classes of functions by removing useless and redundant information. Therefore, when the discrimination model concentrates the attention on a

reduced domain of FTS, one can improve the classification accuracy by decrease the overfitting risk at the same time we reduce the computational burden of the classification model.

Without loss of generality, in the sequel we consider a binary classification problem– the method can be extended to multi class problems straightforwardly– involving two stochastic processes $X_Y(t)$ with classes given by $Y \in \{1, 2\}$. We assume that the domain of the processes can be partitioned in two disjoint non–empty sets $T = \Theta \cup \Theta^C$, such that:

$$P(\omega \in \Omega : X_1(t) = X_2(t)) = 1 \text{ for all } t \in \Theta^C, \tag{6.4}$$

that is $X_1(t) \overset{e.e.}{=} X_2(t)$–the two processes are stochastically equivalent– in $t \in \Theta^C$. Then by removing the domain of the processes where $X_1(t) \overset{e.e.}{=} X_2(t)$ we should at least maintain the performance of any reasonable classification method $h \in \mathcal{C}$; in other words

$$P(h_\Theta(X_Y) = y \,|\, Y = y) \geq P(h(X_Y) = y \,|\, Y = y),$$

for $y = 1, 2$, where h_Θ is the classification method that takes as an input the FTS data on the restricted domain Θ; and this of course needs to be true for any h on a wide class of classification models denoted by $\mathcal{C}$. The goal of next section is to present a divergence curve, and related concepts, in order to estimate the compact set Θ independently of h in $\mathcal{C}$.

6.2 Methodology

6.2.1 Divergence curve: Extending the KL divergence

The *Kullback–Leibler* divergence (KL) is a non–symmetric function that account for the difference between two probability distributions. If we consider two continuous stochastic processes $X_1(t)$ and $X_2(t)$, and fix $t = t_0 \in T$, then the KL divergence is defined as

$$\text{KL}(X_1(t_0)||X_2(t_0)) = \int_{-\infty}^{+\infty} P_{1,t_0}(z) \log\Big(\frac{P_{1,t_0}(z)}{P_{2,t_0}(z)}\Big) dz,$$

where P_{1,t_0} and P_{2,t_0} are the probability functions of the random variables $X_1(t_0)$ and $X_2(t_0)$. Under the condition introduced in Equation 6.4 it holds that $\text{KL}(X_1(t)||X_2(t)) = 0$ for $t \notin \Theta$, and conversely $\text{KL}(X_1(t)||X_2(t)) > 0$ for $t \in \Theta$. This pave the way to introduce the concept of Divergence Curve (KL_C) between the two processes as follows:

Definition 6.1 (Divergence Curve)**.**

$$\mathrm{KL}_C(X_1||X_2) \equiv \{t, \mathrm{KL}(X_1(t)||X_2(t)) \,|\, t \in T\} \tag{6.5}$$

Example 6.1. [*Divergence curve for Gaussian Processes*] Let X_Y, for $Y \in \{1, 2\}$ be two Gaussian processes with mean functions $\mu_Y(t) = \mathbb{E}(X_Y(t))$ and variances $\sigma_Y^2(t) = \mathbb{E}[(X_Y(t) - \mu_Y(t))^2]$, then KL_C can be written as follows:

$$\left\{t, \log\left(\frac{\sigma_2^2(t)}{\sigma_1^2(t)}\right) + \left(\frac{\sigma_1^2(t) - \sigma_2^2(t) + (\mu_1(t) - \mu_2(t))^2}{2\sigma_2^2(t)}\right)\right\}, \tag{6.6}$$

• • •

for all $t \in T$. When $\mu_1(t) = \mu_2(t)$ and $\sigma_1^2(t) = \sigma_2^2(t)$, then the divergence at t equals zero, conversely as the two mean and/or variance functions becomes far apart, the divergence at moment t increases. Assuming standard regularity conditions in the processes under study, the discrimination curve is continuous and well behaved on T –as in the case of Gaussian Processes above–.

6.2.2 A scale–location model

In line with the illustration in Example 6.1, in our numerical experiment we consider $X_1(t)$ and $X_2(t)$, to be two Gaussian processes:

$$\begin{aligned} X_1(t) &\sim \mathcal{GP}(\mu(t); \sigma(s,t)), \\ X_2(t) &\sim \mathcal{GP}(\mu(t)(1 + g_\Theta(t)); \sigma(s,t)), \end{aligned} \tag{6.7}$$

where $\mu(t) = \sin(\pi t)$ and $\sigma(s,t) = \exp^{-15(s-t)^2}$, are the mean of X_1 and the common variance–covariance functions of both processes defined on $T = [0, 1]$. The function $g_\Theta(t) \sim \mathcal{N}(0.5, 0.25)$ for $t \in \Theta = [0.1, 0.9] \subset T$ and $g_\Theta(t) = 0$ for $t \notin \Theta$ is playing the role of a *sacale–location parameter* in the model. In Figure 6.1, we show 100 realizations of X_1 and X_2 in solid blue (—) and red (—) lines respectively; the mean functions $\mu_1(t) = \mu(t)$ and $\mu_2(t) = \mu(t)(1 + g_\Theta(t))$ are represented in (–) and (–) lines respectively. The divergence curve is shown in (–) line, and its estimation counterpart with the (·····) line. In this chapter we only consider scale–location differences between the mean of the processes, but this model can be extended to the case where the differences also occur between the covariance functions by introducing another g_Θ–type function accordingly.

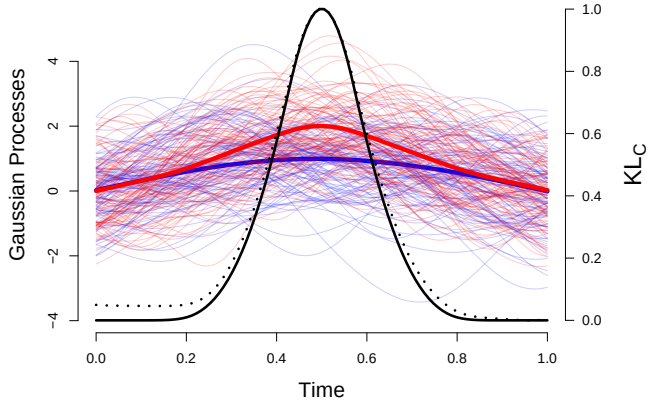


Figure 6.1: Left axis: 50 Realizations of $X_1(t)$ and $X_2(t)$ in solid blue (–) and red (–) lines; mean functions in (━) and (━) respectively. Right axis: KL_C in (━) and its estimated counterpart in (⋯⋯).

6.2.3 Estimating the divergence curve

Given two random samples of functional data $\{X_{Y,j}(t)\}_{j=1}^{n_Y}$ for $Y = 1, 2$, with $n = n_1 + n_2$, in this section we briefly introduce estimators of the divergence curve. As stated in Section 6.1.1, for each realization of the functional time series $x_{Y,j}(t)$ we have a functional estimator $\tilde{x}_{Y,j}(t)$. In this work we use a RKHS approach, but any other alternative basis can be used equivalently. Since each functional datum $\tilde{x}_{Y,j}(t)$ is a continuous object, we can choose an arbitrary grid of time points in T, say $\widetilde{T} = \{t_1, \ldots, t_m\} \subset T$ and compute the in-sample or empirical KL curve using pointwise density estimators for both distributions $P_{1,t}$ and $P_{2,t}$ for any $t \in \widetilde{T}$. For computational reasons and without any considerable impact on the estimation results, in this work we consider $\widetilde{T}$ as the set of discrete time points where the functions are recorded. The discrete version of the KL divergence curve is defined then:

$$\widehat{\mathrm{KL}}_C(X_1||X_2) \equiv \{t, \widehat{\mathrm{KL}}(X_1(t)||X_2(t)) \,|\, t \in \widetilde{T}\},$$

where $\widehat{\mathrm{KL}}$ is the estimation of the KL divergence at t using standard nonparametric Nadaraya–Watson density estimators Wand and Jones (1994) to compute $\widehat{P}_{1,t}$ and $\widehat{P}_{2,t}$. We propose to use non–parametric estimators since they are useful in the setting of arbitrary-shape distributions coming from complex real-world data problems. We do not explore alternative density estimation methods here since it is out of the scope of this chapter, nonetheless, other density estimation methods can be considered, for instance the classical parametric Maximum–Likelihood estimators particularly ubiquitous for Gaussian–Processes.

To make results more interpretable, we embed $\widehat{\mathrm{KL}}_C$ into the interval $[0, 1]$, by using the constants $M = \max_{t\in\widetilde{T}} \widehat{\mathrm{KL}}_C(t)$ and $m = \min_{t\in\widetilde{T}} \widehat{\mathrm{KL}}_C(t)$, as follows

$$\widehat{\mathrm{KL}}_C^S(t) = \frac{\widehat{\mathrm{KL}}_C(t) - m}{M - m}, \tag{6.8}$$

therefore when $\widehat{\mathrm{KL}}_C^S(t) \to 0$ implies that at point $t \in \widetilde{T}$ of the domain, empirically the two classes of functions present a low divergence evidencing that the distribution of X_1 and X_2 at t is similar. On the other hand, when $\widehat{\mathrm{KL}}_C^S(t) \to 1$ the two classes of functions present a high empirical divergence at t, evidencing that the distribution of X_1 and X_2 at t is different. The intervals in the domain where the estimated divergence is high, that is $\{t \in T|\widehat{\mathrm{KL}}_C^S(t) \geq \delta\}$ for some suitably chosen parameter δ, are the one of interest for discrimination purposes. In order to estimate those intervals from data, an additional smoothing can be done using for instance polynomial regression techniques to smooth $\widehat{\mathrm{KL}}_C^S$.

6.2.4 Alternative approach: Common–support proximity curve

To tackle the computational burden when estimating KL_C^S over a large grid $\widetilde{T}$, we also propose a less–informative but computationally cheaper approach as follows: For all $t \in \widetilde{T}$ instead of estimate pointwise the KL divergence using $\widehat{P}_{1,t}$ and $\widehat{P}_{2,t}$, we resort on the computation of the *common–support* between the two distributions by defining first the pointwise in–sample *common* support:

$$A_t = [\max\left(\tilde{x}_{1,[1]}(t), \tilde{x}_{2,[1]}(t)\right), \min\left(\tilde{x}_{1,[n_1]}(t), \tilde{x}_{2,[n_2]}(t)\right)],$$

for $t \in \widetilde{T}$, where $\tilde{x}_{Y,[j]}(t)$ is the in–sample j–th order statistics of the random variable $X_Y(t)$ at $t \in \widetilde{T}$ for $Y = \{1, 2\}$. Let λ be the Lebesgue measure on the real line, and consider $I_t = [\min(\tilde{x}_{1,[1]}(t), \tilde{x}_{2,[1]}(t)), \max(\tilde{x}_{1,[n_1]}(t), \tilde{x}_{2,[n_2]}(t))]$, the following ratio measure the proportion of support that is *common* to the two classes of functions.

$$S(t) = \frac{\lambda(A_t)}{\lambda(I_t)} \text{ for } t \in \widetilde{T}. \tag{6.9}$$

The definition of a common–support curve follow straightforwardly.

Definition 6.2 (Common–Support Curve)**.**

$$\mathrm{S}_C(X_1, X_2) \equiv \{t, S(t)\} \text{ for } t \in \widetilde{T}. \tag{6.10}$$

As well as the KL_C, S_C helps to identify in which subsets of T the two classes of FTS differ more; notice however that S_C is not defined at a population level. By definition S_C is bounded in $[0, 1]$. As $\mathrm{S}_C(t) \to 1$, empirically the conditional distributions of both processes on t shares a wide part of its supports, indicating that the divergence between the two processes should be small. This is a conjecture and not a fact, since we do not estimate the distribution. Conversely, when $\mathrm{S}_C(t) \to 0$, the empirically evidence is in favor of a high divergence value at t.

6.2.5 Using KL_C^S and S_C for the domain selection

In Figure 6.2 (upper panels) on top we shown data simulated from the Gaussian model introduced in Equation 6.7. As can be seen in Figure 6.2 (bottom–left), when we compare the empirical distribution of both processes at $t = 0.05$–i.e. $\widehat{P}_{Y=1,t=0.05}$ against $\widehat{P}_{Y=2,t=0.05}$– they are similar, consequently the support shared by $\widehat{P}_{1,t=0.05}$ and $\widehat{P}_{2,t=0.05}$ is wide; and this correspond in Figure 6.2 at a high value of S_C and a small value of $\widehat{\mathrm{KL}}_C$. Conversely, in Figure 6.2 (bottom–right) we show the empirical distribution of both processes conditional at $t = 0.50$, and those distributions are different. In this case, the divergence between the two distributions is high and the support shared by $\widehat{P}_{1,t=0.50}$ and $\widehat{P}_{2,t=0.50}$ is small; and this corresponds, in Figure 6.2 (upper–right) to a small value of S_C and a high value of $\widehat{\mathrm{KL}}_C$. As said before, the intervals of interest in the context of domain selection are those where $\mathrm{KL}_C^S \geq \overline{\delta}$ or equivalently $\mathrm{S}_C \leq \underline{\delta}$, for suitable chosen threshold values $\overline{\delta}$ and $\underline{\delta}$ respectively. Next Theorem –its proof follows straightforwardly– provides a necessary condition for the existence of such thresholds.

Theorem 6.1. *Let $m \leq f(t) \leq M$ be a semi–continuous and non–constant function on T, then for any value $\overline{\delta} \leq M$, there exist at least one interval $(a, b) \in T$, with $a \leq b$ such that $f(t) \geq \overline{\delta}$ for $t \in (a, b)$. Moreover, for any value $\underline{\delta} \geq m$, there exist at least one interval $(c, d) \in T$, with $c \leq d$ such that $f(t) \leq \underline{\delta}$ for $t \in (c, d)$*

Since by hypothesis we consider enough regularity conditions on the processes under consideration such that KL_C and S_C are semi–continuous and bounded functions on a compact domain T, and there exist a subset $(\theta_1, \theta_2) \in T$ such that $X_1(t) \overset{e.e.}{\neq} X_2(t)$, then the conditions of Theorem 6.1 holds for functions KL_C^S and S_C. For identification reasons, we also assume that KL_C^S is uniquely maximized and S_C is uniquely minimized at some point t in T. This last assumption play no role on the domain selection problem in the practice, since in case of multiple local maxima on KL_C^S or minima in S_C, it can always be assumed that the selected domain is the one that results from the open cover of all the intervals determined by the respective curve.

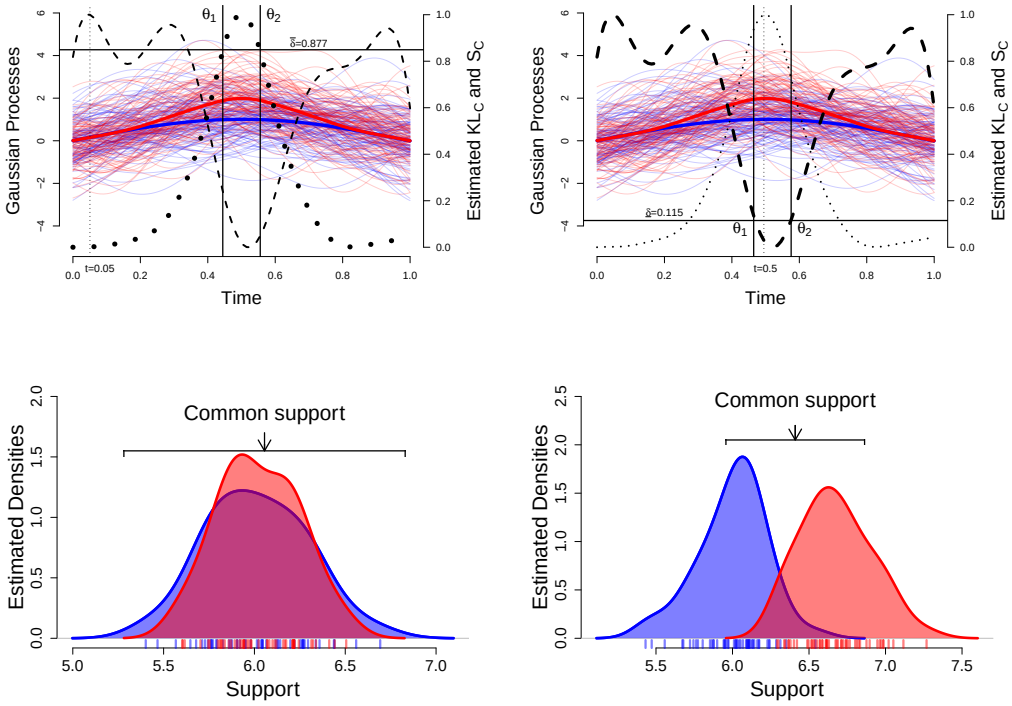


Figure 6.2: Using $\widehat{\mathrm{KL}}_C$ and S_C as a domain selection tool: An illustration. Upper panels: $\widehat{\mathrm{KL}}_C$ and S_C in doted and dashed black lines respectively. Bottom panels: Estimated $\widehat{P}_{1,t}$ and $\widehat{P}_{2,t}$ at $t = 0.05$ and $t = 0.5$ respectively.

In order to unify the estimation of suitable constant $\overline{\delta}$ and $\underline{\delta}$, in this chapter we consider the relationship between the empirical distribution function of KL_C^S and S_C^S and the thresholds as follows:

$$\widehat{P}\big(\widehat{\mathrm{KL}}_C^{\,S} \geq \overline{\delta}\big) = 1 - \nu \text{ and } \widehat{P}\big(\mathrm{S}_C \leq \underline{\delta}\big) = 1 - \nu,$$

so that for for any $\nu \in [0, 1)$, there always exist a unique pair $(\underline{\delta}, \overline{\delta})_\nu$. Notice that for $\nu = 0$, then $\underline{\delta} = \sup \mathrm{S}_C(t)$ and $\overline{\delta} = \inf \widehat{\mathrm{KL}}_C^{\,S}(t)$, which implies the selection of the whole domain. Conversely, as $\nu \to 1$, we shrink the selected domain.

In Figure **??** panels (a) and (b) we show the selected domain (θ_1, θ_2) corresponding to $\nu = 0.9$ ($\overline{\delta} = 0.877$ and $\underline{\delta} = 0.115$) when using $\widehat{\mathrm{KL}}_C^{\,S}$ and S_C respectively. In the practice, to learn a suitable value of ν we regard to standard cross–validation techniques in the context of supervised classification problems. Even more, we also recommend to conduct a sensitivity analysis to quantify the robustness of the proposed domain selection method when tuning the hyper parameter ν.

6.3 Experimental section

6.3.1 Simualtion study

The first natural question is on addressing the estimation performance of the divergence curve. To this aim, in Table 6.1 we presents the results of a Monte Carlo experiment carried out with 1000 instances of the Gaussian model introduced in Equation 6.7 for different samples sizes, namely $n_Y = \{10, 50, 100, 250, 500, 1000, 5000\}$ for $Y = \{1, 2\}$. We report the mean square error (MSE) defined as

$$\text{MSE}_n = \int_0^1 (\text{KL}_C(t) - \widehat{\text{KL}}_C(t))^2 \, dt,$$

as can be seen, the estimated divergence curve converge at a high rate to the true curve as the sample size increases.

Table 6.1: Average MSE (avg). Standard errors (sd.) are reported in parenthesis.

n_Y:	10	50	100	250	500	1000	5000
avg.	0.065	0.026	0.014	0.011	0.010	0.001	0.000
sd.	(0.0879)	(0.0118)	(0.0058)	(0.0026)	(0.0010)	(0.0009)	(0.0001)

The second goal on this numerical experiment, is to address how the domain selection affects the performance of a wide class of classification models in $\mathcal{C}$. We select as representatives the following well known discrimination methods in the functional data context:

Suppor Vector Machines (SVM): We consider the representation coefficients of each FTS as the imput variables following the methodology proposed by Muñoz and González (2010). The method is implemented using the `e1071` R–package. We use the Gaussian Kernel function and tune its parameter by cross–validation.

K–Nearest Neighbors (KNN): This classification algorithm is applied considering as input variables the representation coefficients $\alpha' s$.

FunClust (FC) Bouveyron and Jacques (2011) and **FunHDDC** (FHDDC) Jacques and Preda (2013): These two adaptive clustering techniques use the functional principal component analysis (FPCA) scores to represent the functional time series. Both modelization scheemes assume the functional principal components follows a Gaussian distribution and use a probabilistic approach to estimate the probabilities by means of the following mixture model:

$$f_{\tilde{X}}^{(q)}(\tilde{x};\Delta) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{q_k} f_{C_{j|Z_k=1}}(c_{jk(\tilde{x})};\lambda_{jk}),$$

where $\Delta = \{\pi_k, \lambda_{1k}, \ldots, \lambda_{q_k k}\}_{k=1}^{K}$ are the parameters of the model (cluster proportions and FPCA variances) and q_k is a truncation parameter for cluster k. The main difference between the *Funclust* and *FunHDDC* is that in the former the FPCA $c_{jk}(x)$ is estimated through an EM algorithm that computes the conditional probabilities of the curves to belong to each cluster and then define the truncation parameter q_k, while in the *FunHDDC* each truncation parameter q_k is fixed to the maximum number of basis functions considered in FPCA. This procedures are already implemented in the R–packages `Funclustering` and `funHDDC`. For further details see Bouveyron and Jacques (2011); Jacques and Preda (2013, 2014).

Maximum depth classifier Li et al. (2012): This model assign the functional datum $\tilde{x}(t)$ the class k if $\mathrm{D}_1(\tilde{x}(t)) \geq \mathrm{D}_2(\tilde{x}(t))$, where D_k is a *depth measure* computed with the in–sample functional data of classes $k = 1, 2$. In this chapter we consider as depth measures: the Fraiman and Muniz depth (FM) Fraiman and Muniz (2001), the h-mode depth (HM) and the Random Projection depth (RP) Cuevas et al. (2007) implemented in the `fda.usc` R–package, Febrero-Bande et al. (2012).

Simulation details: Using the same Gaussian framework introduced in Equation 6.7, with a second Monte–Carlo experimentation we study the performance of the proposed domain selection method applied to the previous list of classifiers. We keep the number of curves sampled on each classes fixed to $n_Y = 100$, for $Y = \{1, 2\}$, where 50% of the curves are considered as training samples to learn the effective domain for different threshold values ν, and the remaining 50% to test the out–of–sample classification performance of the previous models over the the learned domain. We replicate the data generation process 1000 times–the first instance of these simulations is depicted in Figure 6.1– to estimate the average out–of–sample error and its variability for each method and threshold value ν.

In Tables 6.2 and 6.3 we present the average out–of–sample classification errors, for different methods –in rows– and for different threshold values ν –in columns– (i.e., we are using different domains for each column), when using KL_C and S_C, respectively. In both tables, it can be seen that as we constrain the effective domain to regions where the divergence between the two classes of functions increases (or equivalently, where the common support reduces), all the classification methods improve the performance since

Table 6.2: Monte-Carlo study: Estimated average out–of–sample errors for different classification methods when using the domain selected by KL_C for different threshold ν_δ values –in columns–. Standard errors are reported in parenthesis.

Method (boost %)	$\nu = 0$ (Global)	$\nu = 0.5$	$\nu = 0.8$	$\nu = 0.9$
FM (4.48%)	0,256 (0,029)	0.253 (0,033)	0.251 (0,035)	0.245 (0,033)
HM (11.45%)	0,253 (0,028)	0,242 (0,029)	0,236 (0,030)	0,232 (0,032)
RP (14.1%)	0,259 (0,030)	0,253 (0,030)	0,246 (0,031)	0,237 (0,034)
KNN (+100%)	0,295 (0,029)	0,002 (0,003)	0,000 (0,000)	0,000 (0,000)
SVM (+100%)	0,028 (0,028)	0,001 (0,001)	0,000 (0,000)	0,000 (0,000)
FunClust (34.5%)	0,467 (0,022)	0,458 (0,028)	0,461 (0,028)	0,346 (0,112)
FunHDDC (19.24%)	0,316 (0,018)	0,281 (0,015)	0,270 (0,005)	0,265 (0,000)
Divergence (train)	0.00	0,087	0,506	0,856
θ_1	1	24	42	47
θ_2	100	73	61	56
Selected Domain (train) (%)	100	50	20	10

Table 6.3: Monte-Carlo study: Estimated average out–of–sample errors for different classification methods when using the domain selected by S_C for different threshold ν_δ values –in columns–. Standard errors are reported in parenthesis.

Method (boost %)	$\nu = 0$ (Global)	$\nu = 0.5$	$\nu = 0.8$	$\nu = 0.9$
FM (10.82%)	0,256 (0,029)	0,254 (0,031)	0,239 (0,032)	0,231 (0,032)
HM (11.45%)	0,253 (0,028)	0,245 (0,031)	0,236 (0,029)	0,227 (0,032)
RP (14.09%)	0,259 (0,030)	0,257 (0,032)	0,234 (0,031)	0,227 (0,032)
KNN (+100%)	0,295 (0,029)	0,101 (0,021)	0,000 (0,001)	0,006 (0,005)
SVM (+100%)	0,028 (0,028)	0,057 (0,016)	0,000 (0,000)	0,006 (0,005)
FunClust (32.67%)	0,467 (0,022)	0,450 (0,028)	0,459 (0,027)	0,352 (0,107)
FunHDDC (21.53%)	0,316 (0,018)	0,312 (0,009)	0,290 (0,001)	0,260 (0,001)
Common–Support (train)	0.948	0,806	0,716	0,642
θ_1	1	1	24	50
θ_2	100	64	61	59
Selected Domain (train) (%)	100	64	38	10

we are able to reduce the estimated miss-classification error rate (by reducing type–I and type–II errors). The estimated on average out–of–sample error reduction of each method is presented in parenthesis on the first column of each table, we can see that all the methods show incredible high increases of performance, some of them more than 100% and none of them below 10%, when we compare the performance staring with all the domain (the standard approach) with respect to the best case when constraining

the domain (most of the cases to only 10% of the original interval). As an associated advantage of the domain selection methodology, we must mention the **reduction in computational costs** associated with each classification method. For the case of this simulation, we are allowed to reduce the data storage by approximately 90%–plus the computational time gain associated with the calibration of the discrimination model with a reduced domain–.

6.3.2 Real data examples

To illustrate the procedure with real problem in science and technology, we consider two different functional time series data sets taken from the UEA & UCR Time Series Classification Repository. In all the cases, to evaluate the predictive performance of each classification method, we consider a training sample to learn the relevant domain of the functions, and the test sample to evaluate the classification performance of each method.

Chinatown Pedestrian Curves. The Pedestrian Counting System dataset of the City of Melbourne, Australia is an automated counting system that helps to understand the dynamics and patterns of the pedestrian activity within the city, and therefore make better decisions at an urban planning level. This data set consist of pedestrian count in the Chinatown-Swanston St North, for the 12 months of 2017. Classes distinguish between working days (class 2) and non–working days (class 1). The trainging and testing sample size of this data set is 20 and 356 days respectively sampled at each hour of the day (24 points). The inference excersise here is to classify wokring from non–working days, –see UEA & UCR and Merlburne Pedestrian Counting System web–site for further details–.

Even though the means of the working and non–workings days present difference along the entire domain of the curves, the discriminations metrics show that the main difference between the two classes of days in the pedestrian counts are given at the first hours of the day (01:00–05:00) –see Figure 6.3.2–. In particular this sub–domain helps to better discriminate the classes and improve the out–of–the–sample classification results, see 6.4. Besides the boosting of the classficiation problem, the identification of the subdomain is key to understand the pedestrian behaviour at Chinatown-Swanston St North which or any corner of the city –if the data is available–, being able to be an useful input for the development of an urban plan.

Italy Power Demand. This data set was firstly considered by Keogh et al. (2006) and

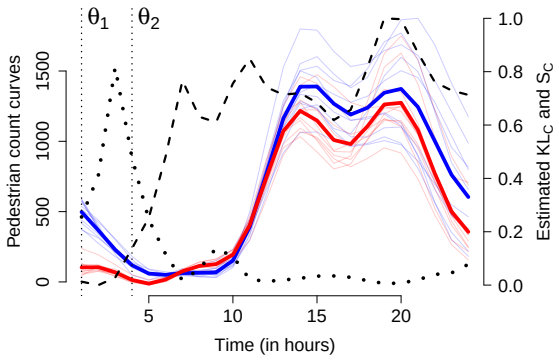


Figure 6.3: Chinatown Pedestrian Curves: In blue (–) non–working days, in red (–) working days, in)–) and (–) the respective means. The (⋯⋯) and (- - -) lines corresponds to estimated KL_C^S and S_C respectively. The domain selected correspond to $\nu = 0.86$: $\theta_1 = 1$ and $\theta_2 = 4$ in vertical black lines.

consists of intra–day power demand curves in Italy. Interesting data analysis may be conducted with smart meter type of data, for instance the classification of targeted group of clients according to its demand consumption, but for this chapter we maintain the classes identified by the mentioned authors, namely: The electricity demand corresponding to winter months (October to March) and the ones that belongs to summer months (April to September); therefore, the classification task is to distinguish winter days from summer days. In Figure 6.3.2 we shown the training FTS data –constituted by 67 days (curves)– , in color red the curves corresponding to summer days and in blue those from winter days.

Using the domain selection approach introduced in this chapter, we are able to improve the classification performance of many discrimination methods, at the same time we can also learn at what time during the day the consumption pattern differ the most between the seasons. To this aim, the training set is used to learn the domain where the consumption pattern differ the most between the seasons. Then, using the testing set–contains 1029 days (curves)– we estimate the out–of–sample classification error of all discrimination methods introduced earlier. Table 6.5 show the testing errors for different classification methods and the selected domains. The results indicate that the classification error rate decreases as a particular sub interval of the whole domain is considered in the classification process. In particular in Table 6.5 for the *FunHDDC* method with a level of divergence (δ) of 0.65, which is accumulated at the quantile 0.95, the classification error is reduced in average to 3%. In this case the estimated domain is $\Theta = \{19, 20\}$.

Table 6.4: Chinatown Pedestrian Curves. Testing errors, Optimal Selected Domain (θ_1, θ_2), usinge the KL_C^S and the S_C divergence.

Method	Metric	Global	KL_C	S_C
	Test Error	0.159	0.026	0.020
	θ_1	1	3	1
FM	θ_2	24	4	3
	δ	0	0.468	0.028
	ν	0	0.95	0.91
	DS(%)	100	8.33	12.50
	Test Error	0.064	0.023	0.023
	θ_1	1	1	1
HM	θ_2	24	5	5
	δ	0	0.231	0.066
	ν	0	0.82	0.82
	DS(%)	100	20.83	20.83
	Test Error	0.252	0.032	0.032
	θ_1	1	1	1
RP	θ_2	24	10	9
	δ	0	0.103	0.407
	ν	0	0.69	0.73
	DS(%)	100	41.67	37.50
	Test Error	0.241	0.151	0.168
	θ_1	1	1	1
KNN	θ_2	24	5	2
	δ	0	0.231	0.012
	ν	0	0.82	0.95
	DS(%)	100	20.83	8.33
	Test Error	0.177	0.023	0.188
	θ_1	1	3	1
SVM	θ_2	24	4	17
	δ	0	0.468	0.459
	ν	0	0.95	0.60
	DS(%)	100	8.33	70.83
	Test Error	0.255	0.154	0.165
	θ_1	1	2	1
FunClust	θ_2	24	4	9
	δ	0	0.401	0.407
	ν	0	0.91	0.73
	DS(%)	100	12.50	37.50
	Test Error	0.243	0.026	0.026
	θ_1	1	1	1
FunHDDC	θ_2	24	10	9
	δ	0	0.103	0.407
	ν	0	0.69	0.73
	DS(%)	100	41.67	37.50

ECG data set. The ECG data set was structured by Olszewski (2001) in his doctoral disertation. Each curve represent the cardiac electrical activity recorded during one hearbeat. The two classes preesent normal patients and patients with Myocardial Infarction (Ischemia). The trainging and testing sample size of this data set is 100 signlas sampled at 96 points. In this case the empirical excercise has a biometric implication and is not only to distinguish healthy from non–healthy patiens, also involes the excer-

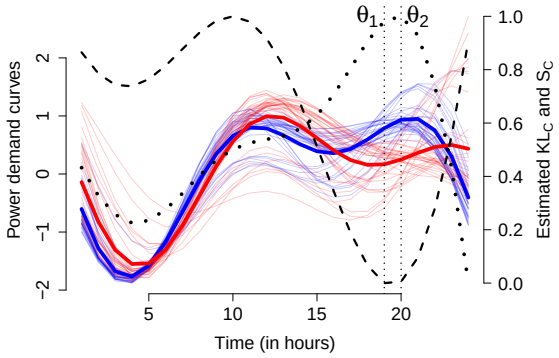


Figure 6.4: Power demand curves: In blue (–) winter days, in red (–) summer days, in (–) and (–) the respective means. The (·····) and (- - -) lines corresponds to estimated KL_C^S and S_C respectively. The domain selected correspond to $\nu = 0.95$: $\theta_1 = 19$ and $\theta_2 = 20$ in vertical black lines.

cise of indentify in which part of the heartbeat do the difference between the classes of patients is larger. For further details see Olszewski (2001).

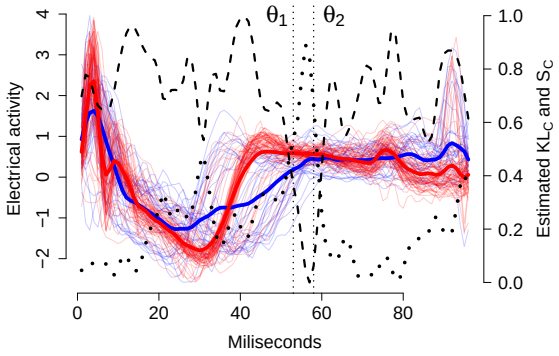


Figure 6.5: Electrocardiograms data set : In blue (–) normal myocardial activity, in red (–) patients with Ischemia, in (–) and (–) the respective means. The (·····) and (- - -) lines corresponds to estimated KL_C^S and S_C respectively. The domain selected correspond to $\nu = 0.94$: $\theta_1 = 53$ and $\theta_2 = 58$ in vertical black lines.

The heart electrical activity experiment throw really interesting results. Firstly, analyzing Figure 6.3.2 it can be appreciated that, although the means of the electrical activiy for healthy patients and patients with Ischemia present the major differences whitin the 20–60 miliseconds of the hearbeat, both the KL_C^S and the S_C suggest that the largest divergence (or lowest common–support) is given at the percentile $94^{\underline{th}}$ of the metric, generating a subdomain located between the 53–58 miliseconds of the heartbeat. This reduce the domain from 96 miliseconds to only 6 (approximately 99% less); a reduc-

Table 6.5: Italy Power Demand Data Set. Testing errors, Optimal Selected Domain (θ_1, θ_2), usinge the KL_C^S and the S_C divergence.

Method	Metric	Global	KL_C	S_C
	Test Error	0.144	0.160	0.164
	θ_1	1	6	3
FM	θ_2	24	21	21
	δ	0	0.142	0.489
	ν	0	0.65	0.56
	DS(%)	100	66.67	79.17
	Test Error	0.045	0.045	0.045
	θ_1	1	1	1
HM	θ_2	24	24	24
	δ	0	0.062	0.588
	ν	0	0.21	0.39
	DS(%)	100	100	100
	Test Error	0.108	0.063	0.080
	θ_1	1	19	21
RP	θ_2	24	21	21
	δ	0	0.650	0.561
	ν	0	0.95	0.43
	DS(%)	100	8.33	87.50
	Test Error	0.311	0.131	0.248
	θ_1	1	11	4
KNN	θ_2	24	21	20
	δ	0	0.303	0.438
	ν	0	0.82	0.69
	DS(%)	100	45.83	70.83
	Test Error	0.273	0.082	0.178
	θ_1	1	6	4
SVM	θ_2	24	21	21
	δ	0	0.142	0.477
	ν	0	0.65	0.60
	DS(%)	100	66.67	75
	Test Error	0.297	0.446	0.465
	θ_1	1	19	4
FunClust	θ_2	24	20	21
	δ	0	0.650	0.477
	ν	0	0.95	0.60
	DS(%)	100	8.333	75
	Test Error	0.060	0.038	0.072
	θ_1	1	19	6
FunHDDC	θ_2	24	20	20
	δ	0	0.650	0.325
	ν	0	0.95	0.95
	DS(%)	100	8.33	62.5

tion that also involve the information included in the electrocardiograms as well. On the other hand, the out–of–the–sample classification results show an average increase in the performance of identifying healthy patients and patients with ischemic cardiac disease when the classification method consider the optimal selected domain for each discrimination metric –see Table 6.6.

Table 6.6: Electrocardiogram data set. Testing errors, Optimal Selected Domain (θ_1, θ_2), usinge the KL_C^S and the S_C divergence.

Method	Metric	Global	KL_C	S_C
FM	Test Error	0.280	0.240	0.230
	θ_1	1	31	30
	θ_2	96	59	87
	δ	0	0.445	0.692
	ν	0	0.91	0.83
	DS(%)	100	30.21	60.41
HM	Test Error	0.210	0.210	0.210
	θ_1	1	54	54
	θ_2	96	58	59
	δ	0	0.553	0.558
	ν	0	0.95	0.94
	DS(%)	100	5.21	6.25
RP	Test Error	0.230	0.210	0.210
	θ_1	1	30	30
	θ_2	96	59	87
	δ	0	0.421	0.692
	ν	0	0.90	0.83
	DS(%)	100	31.25	60.42
KNN	Test Error	0.310	0.390	0.340
	θ_1	1	29	55
	θ_2	96	96	58
	δ	0	0.339	0.488
	ν	0	0.84	0.96
	DS(%)	100	70.83	4.16
SVM	Test Error	0.270	0.220	0.210
	θ_1	1	55	53
	θ_2	96	57	60
	δ	0	0.680	0.619
	ν	0	0.97	0.92
	DS(%)	100	3.12	8.33
FunClust	Test Error	0.350	0.300	0.280
	θ_1	1	31	31
	θ_2	96	59	86
	δ	0	0.445	0.668
	ν	0	0.91	0.89
	DS(%)	100	30.21	58.33
FunHDDC	Test Error	0.243	0.026	0.026
	θ_1	1	1	1
	θ_2	24	10	9
	δ	0	0.103	0.407
	ν	0	0.69	0.73
	DS(%)	100	41.67	37.50

Satellite image classification. This data set was created by the authors in Tan et al. (2017). The data consist of high definition images taken by FORMOSAT-2 satellite, and each pixels on the image corresponds to a specific geographic area on Earth of 64 square meter. For each pixel, the Normalized Difference Vegetation Index (NDVI) is recorded at 46 different moments during the year, and the main goal of this data is on the identification of land uses trough the evolution of NDVI index during time. Originally, the

data set contains 24 classes of curves but for space reason in this empirical example we conducted a one–against–one classification problem between classes corresponding to softwood, poplars, sorghum and barley. Training and testing size are 1200 curves (300 on each class) and 2800 (700 on each class) respectively. An illustration of the data set is presented in Figure 6.6. For further details on this data set see Tan et al. (2017).

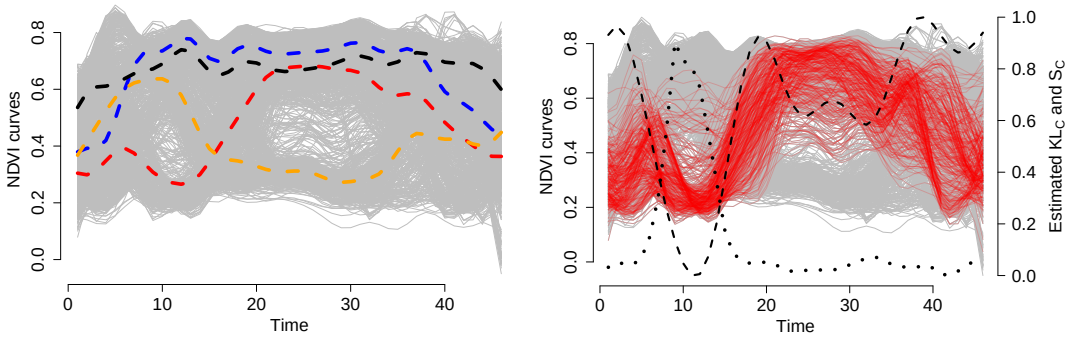


Figure 6.6: Training set and domain selection metrics.Training set and domain selection metrics. Left panel: NDVI curves–all classes– and the mean functions corresponding to Softwood (- - -), Poplars (- - -), Sorghum (- - -) and Barley (- - -). Right panel: NDVI curves of poplars (in —) vs. the rest (—) and the estimated KL_C^S (·····) the and S_C (- - -).

For this multiclass experiment we developed a *one–against-all* scheme. The results presented in Table 6.7 show that for all the exercises exists at least one sub-interval of the domain where the classification method improve the classification results in comparison with the whole domain approach. This empirical real data example show that guided by the KL divergence, one can boost the classification error rates. This means that an important asset in time series classification, is the analysis of the temporal dynamics of the divergence, or in the case of the common–support criteria, the "proximity", between the classes of functions.

6.4 Chapter summary

In this chapter we propose a novel method for *domain selection* with functional time series data. We introduce the concept of *divergence curve* and its estimator counterpart, and the *common support curve*, a computationally cheaper but method for droping out redundant information in the context of supervised classification problems.

We show with the aid of a numerical simulations study that the proposed methodology has associated improvements in terms of classification performance reducing, at

Table 6.7: NDVI Data Set. Test errors and selected domain $\Theta = \{\theta_1, \theta_2\}$. Domain Selection metric: KL_C^S. Classfication methods: Depth Classifiers FM and RP.

Method	Metric	Fraiman & Muniz (FM)			Random Projections (RP)		
		Global	KL_C	S_C	Global	KL_C	S_C
Softwood	Test Error	0.130	0.130	0.130	0.102	0.108	0.095
	θ_1	1	1	1	1	1	1
	θ_2	46	46	46	46	46	46
	δ	0	0.072	0.836	0	0.072	0.836
	ν	0	0.06	0	0	0.06	0
	DS (%)	100	100	100	100	100	100
Sorghum	Test Error	0.061	0.029	0.05	0.034	0.024	0.027
	θ_1	1	8	7	1	5	6
	θ_2	46	13	33	46	33	36
	δ	0	0.477	0.587	0	0.069	0.721
	ν	0	0.88	0.95	0	0.73	0.46
	DS (%)	100	13.04	58.69	100	63.04	67.39
Barley	Test Error	0.051	0.019	0.018	0.026	0.017	0.017
	θ_1	1	21	21	1	21	21
	θ_2	46	34	35	46	34	34
	δ	0	0.582	0.464	0	0.582	0.446
	ν	0	0.71	0.73	0	0.71	0.75
	DS (%)	100	30.43	32.60	100	30.43	30.43
Poplars	Test Error	0.146	0.139	0.147	0.077	0.078	0.079
	θ_1	1	4	5	1	1	2
	θ_2	46	20	19	46	45	44
	δ	0	0.485	0.346	0	0.170	0.461
	ν	0	0.77	0.84	0	0.28	0.11
	DS (%)	100	36.95	32.60	100	97.82	93.47

the same time and without no costs in terms of classification performance, the computational burden for several functional classification methods.

In the experimental section we conduct the analysis of several functional time series data sets and the empirical results show consistent improvements in supervised classification problems when the effective domain is learned in a first round of the problem.

Acknowledgment

I would like to acknowledge the use of the UEA & UCR Time Series Classification Repository (`http://timeseriesclassification.com/index.php`).

Chapter 7

Conclusions and future work

7.1 Conclusions of the thesis

In this thesis we propose several sophisticated methodologies to tackle problems of statistical inference for functional data. In particular our aim is to solve outlier or anomaly detection, prediction and classification problems in a functional context.

In Chapter 2 we present and discuss the concept of functional data and the implications on statistical inference. In particular we state the theoretical framework that we consider along the whole manuscript. All the statistical learning methods are based on a reproducing kernel Hilbert space model for functional data.

Chapter 3 of this thesis introduce kernel based depth measures for functional data. Two depth measures that induce order into the data were proposed: i) the Kernel Mahalanbobis Depth (KMD), based on the Mahalanobis distance jointly with a robustified version of it and, ii) the Generalized Kernel Depth (GKD) based on a generalization of the Mahalanobis distance via density kernels. We prove that the proposed Generalized Kernel Depth measure fulfil several desirable theoretical properties, in particular the invariance under RKHS bases choice. Morover we show that the Mahalanobid depth and the h–mode depth are particular cases of the Generalized Kernel Depth proposed.

Chapter 4 propose a definition of Entropy for stochastic processes, considering a Reproducing Kernel Hilbert Space model to estimate the Entropy from a random sample of realizations of a stochastic process, namely functional data, and introduce two approaches to estimate minimum entropy sets for functional anomaly detection. we also show the convergence of the parametric Entropy estimation method to the true values

through a montecarlo simulation. Morover the order invariance property is studied for both the parametric and non–parametric approach.

In Chapter 5 we present a new autoregresive Hilbertian model for functional time series. Based on a reproducing kernel Hilbert space framework, the first contribution is to develop a new family of basis functions to estimate the autocorrelation operator Ψ and to predict an entire new function for the whole domain. Throughout several Monte–Carlo studies, we show the performance of the proposed model, in terms of the root mean squared error, against well known prediction methodologies for functional time series.

As a second contribution we tackle the issue of constructing predictive conifidence bands for the point forecast. We present a discussion related to the pointwise and simultaneous inference approaches to construct the predictive bands. Our proposed methodology is based on a model–based bootstrap approach for functional time series, which is an extension of the PRR bootstrap procedure. We theoretically justify our proposal based on the continuity of the integral operator, noticing the advantage of the reproducing kernel Hilbert space framework over other approaches.

In Chapter 6 we propose a novel method for *domain selection* with functional time series data. We introduce the concept of *divergence curve* and its estimator counterpart, and the *common support curve*, a computationally cheaper but equivalent tool necessary for drooping out redundant information in the context of supervised classification problems. We also state the formalism for doing the selection of the effective domain in classification problems and show by simulations that the proposed methodology has associated strong improvements in terms of classification performance reducing, at the same time and without no costs, the computational burden–in terms of memory and time– for several functional classification methods.

7.2 Future research lines

Next we discuss some possible avenues of future research lines.

Specific future research lines related to Chapter 3

Regarding future work, we consider as a priority the study of asymptotic properties of the GKD; in particular to approximate its distribution in order to determine a prob-

abilistic threshold for outlier identification. Another way to go when determining a suitable threshold, is investigating on Bootstrap methods to approximate an empirical optimal cut point. A second natural avenue for future work includes the extension of the concept of depth for functional data without a finite dimensional representation. A third branch of resesearch that we plan to undergo is the study of depth measures over manifolds, –i.e. spheres–.

Specific future research lines related to Chapter 4

A natural extension for future work entails the study of asymptotic properties of the MES_ν estimators. The extension of the proposed method from stochastic process to random fields, useful for several statistical and information science areas, seems straightforward but a wide range of simulations and numerical experiments must be done in order to stress the performance of Entropy methods in comparison to other techniques when dealing with abnormal fields. Another natural avenue for future work entails the study of the connections between Entropy for stochastic process, as formally defined here, and the maximum entropy principle when estimating the governing parameters of Gaussian processes.

Specific future research lines related to Chapter 5

The proposed methodology in this Chapter, has three natural extension. The first one is to study the statistical limiting properties of the functional bootstrap procedure proposed. The second one is the extension of the FA–RKHS model to non–stationary functional time series.

The third one is to extend the methodology to multivariate functional time series framework. One possible avenue of research is to study and analyze the long term relationship and temporal dynamic between two functional data sets. To this aim an estimation of the equilibrium correction operator is needed. This study can be framed in the extension of the Vectore Error Correction models to the functional context. The simplest way to do such extension is to reproduce the two–stage scheme proposed by Engle (1991) as follows,

Consider to non–stationary functional autoregressive processes, X_n and Y_n the first stage involves the estimation of the long run equation

$$X_n = \psi Y_n + \epsilon_n,$$

where X_n and Y_n are a finite sequence of random functions in L^2, $\psi \in L^2$ and ϵ_n is a sequence of iid zero mean errors in L^2, such that $\mathbb{E}\|\epsilon_n\| < \infty$.

Following the Co–Integration procedrure, in the second stage implies to test the functional stationarity of the elements ϵ_n. If they are so, we could study the long run relationship that would be given by,

$$\epsilon_n = X_n + \psi Y_n.$$

Then the Functional Error Correction Model (FECM) is:

$$X_n = \xi(X_{n-1} + \psi Y_{n-1}) + \psi Y_n + \epsilon_n^*,$$

where the operator ξ is the equilibrium correction operator.

Specific future research lines related to Chapter 6

Several natural avenues for future research came after this work. We start mentioning the following statistical aspect of domain selection in the near future:

- Study alternative estimators for the divergence curve, for instance the parametric and semiparametric ones, and more important the conditions on the stochastic process under study in order to get large sample properties for the estimation of KL_C (for instance the consistency).
- Another interesting open question is on the way to conduct inference in the context of this problem; for instance, a methodology to build confidence regions for θ_1 and θ_2, in order to *characterize the uncertainty around domain selection*.
- We give a computationally cheaper version of KL_C curve, namely S_C, but there is no population counterpart for such an object (there is no asymptotic theory to discuss here). It will be desirable to elaborate on alternative–and computationally cheaper– approximations to KL_C with well defined limit in order to attend its large sample properties too.
- The threshold for for both curves are defined using cross–validation. In this context, a future research opportunity is the study of the asymptotic distribution of such a threshold in order to derive a suitable probabilistic method to derive its optimal value and avoid computationally expensive calibrations.

In the other corner, and from a computational view point, we mention the following future steps in terms of research:

- There is a lot of room for improvements in terms of computational performance. We already mention alternatives to KL_C, but also other ways to estimate KL_C that do not rely on the non–parametric estimation of $P_{Y,t}$ will also produce several speed ups in the model.
- Part of the future research is indeed undergoing, we are implementing the R routines in an R package that will be ready soon for the use of the data science and machine learning community.

Appendix A

Appendix to Chapter 3

A.1 Empirical functional median as the deepest curve

Proposition A.1. *The empirical functional median as defined in Def. 3.7 is the curve with highest Band depth and Modified Band depth.*

Proof. Consider a functional data sets $\mathcal{S}_n = \{\tilde{x}_1(t), \ldots, \tilde{x}_n(t)\}$ and $\mathcal{S}_h^{'} = \{\mathcal{S}_n, \tilde{x}_{me}(t)\}$, where $h = n + 1$ and $\tilde{x}_{me}(t)$ is the empirical functional median of the set $\mathcal{S}_n$ given Definition 3.7.
Define the compact interval $I = \{t_i, \ldots, t_m\} \in \mathbb{R}$. Then I can be divided in compact subinvtervals I_j for $j = 1, \ldots, p \leq m$ such that:

(i) For each $t_i \in I_j,\ \tilde{x}_l(t_i) \neq \tilde{x}_k(t_i), \forall\, l, k = 1, \ldots, n$;

(ii) $\bigcup_{j=1}^{p} I_j = I$.

For each subinterval I_j we define the subset of sample curves $\mathcal{S}_h^{'j}$, where $\tilde{x}_l^j(t) \equiv \{(t_i, \tilde{x}_l^j(t_i)) \in I_j \times \mathbb{R}\}_{i=1}^{|I_j|}$, for $l = 1, \ldots, n + 1$. By definition of the Band depth and the Modified Band depth –see def. 3.8 and 3.9– and for each j it holds that:

$$\tilde{x}_{me}^j(t) = arg \max_{\tilde{x}^j(t)} BD(\tilde{x}^j(t), \mathcal{S}_h^{'j}), \text{and} \tag{A.1a}$$

$$\tilde{x}_{me}^j(t) = arg \max_{\tilde{x}^j(t)} MBD(\tilde{x}^j(t), \mathcal{S}_h^{'j}). \tag{A.1b}$$

Therefore, as each subintervals I_j covers I then the result from Eq. A.1a and Eq. A.1b holds for the sample of curves $\mathcal{S}_h^{'}$. □

A.2 Proofs Proposition 3.1

The proof for proposition P.1 and P.2 are proven straightforward by definitions 3.16 and 3.17. To prove P.3 consider $\|\mathbf{z}\| \to \infty$ then it holds that $f_{\mathbb{F}}(\mathbf{z}) \to 0$, which implies that $\text{GKD}(\tilde{x}(t), \mathcal{S}_n, K_F) = \phi_{\mathbb{F}}(\mathbf{z})\phi_{\mathbb{F}}(\widehat{\mathbf{m}}_0) \to 0$.

To prove P.4, consider the affine map $\tau \in \mathcal{T}$ of the form $\tau \circ \tilde{x}(t) = a + b\tilde{x}(t) = \tilde{x}(t)'$. Then from Equation 2.8, it holds that if $\mathbf{z} = (z_1, \dots, z_d)$ represents $\tilde{x}(t)$, then $\mathbf{z}' = (a+bz_1, \dots, a+bz_d)$ is the finite dimensional representation corresponding to $\tilde{x}'(t)$. By definitions 3.16 and 3.17, it holds that

$$\phi_{\mathbb{F}}(\mathbf{z})\phi_{\mathbb{F}}(\widehat{\mathbf{m}}_0) = \phi_{\mathbb{F}}(\mathbf{z}')\phi_{\mathbb{F}}(\widehat{\mathbf{m}}_0') \tag{A.2}$$

where $\widehat{\mathbf{m_0}'}$ is the estimated mode using $\{\mathbf{z}_1' \dots, \mathbf{z}_n'\}$, from where the result follows.

To prove P.5 consider $K_B(r,t) = \sum_{i=1}^{\infty} \lambda_i \phi_i(r)\phi_i(t)$ and $K_G(r,t) = \sum_{i=1}^{\infty} \kappa_i \psi_i(r)\psi_i(t)$ two kernel function –i.e. two different bases– that generate $\mathcal{H} \subset C(T)$, then

$$\text{GKD}(\tilde{x}_d(t)_B, \mathcal{S}_n, K_F) = \text{GKD}(\tilde{x}_d(t)_G, \mathcal{S}_n, K_F),$$

where $\tilde{x}_d(t)_B = \sum_{i=1}^{d} a_i\phi_i(t)$ and $\tilde{x}_d(t)_G = \sum_{i=1}^{d} b_i\psi_i(t)$, for some suitable sequence of coefficients $\{a_i\}_{i=1}^d$ and $\{b_i\}_{i=1}^d$ respectively.

Proof. Let $\{x(t_i), t_i\}_{i=1}^m$ be a discrete realization of the stochastic process $X(\omega, t)$. Consider two kernel functions K_B and K_G, that is two orthonormal bases $B = \{\phi_1, \phi_2, \dots\}$ and $G = \{\psi_1, \psi_2, \dots\}$. Assume that for a fixed d, it holds that $\mathcal{H}_d = \text{Span}\{\phi_1, \dots, \phi_d\} = \text{Span}\{\psi_1, \dots, \psi_d\}$. Then solving the regularization problem stated in Equation 2.3, $\tilde{x}_d(t) \in \mathcal{H}_d \subset \mathcal{H}$ can be expressed in two bases, namely

$$\tilde{x}_d(t)_B = \sum_{i=1}^{m}\sum_{j=1}^{d} \alpha_i\lambda_j\phi_j(t)\phi_j(t_i), \text{ and} \tag{A.3a}$$

$$\tilde{x}_d(t)_G = \sum_{i=1}^{m}\sum_{l=1}^{d} \beta_i\kappa_l\psi_l(t)\psi_l(t_i), \tag{A.3b}$$

then each ϕ_j can be expressed in terms of $\{\psi_l\}_{l\geq 1}^d$–and conversely–, as follows

$$\phi_j(t) = \sum_{i=1}^{m}\sum_{l=1}^{d} \gamma_i\psi_l(t)\psi_l(t_i) \tag{A.4}$$

Then combining Equations A.3a withA.4, we have

$$\tilde{x}_d(t)_B = \sum_{i=1}^{m}\sum_{l=1}^{d} \alpha_i \lambda_j \gamma_i \psi_l(t)\psi_l(t_i). \tag{A.5}$$

Therefore,

$$\tilde{x}_d(t)_B = \sum_{i=1}^{m}\sum_{l=1}^{d} \delta_{ij} \psi_l(t)\psi_l(t_i). \tag{A.6}$$

where $\delta_{ij} = \alpha_i\lambda_j\gamma_i$ for $i = 1, \ldots, m$ and $j = 1 \ldots, d$. Therefore we can represent $\tilde{x}_d(t)_B$ in the subspace generated by the $\mathrm{Span}(G)$ by means of a change of bases –an affine transformation–, hence by P4 $\mathrm{GKD}(\tilde{x}_d(t)_B, \mathcal{S}_n, K_F) = \mathrm{GKD}(\tilde{x}_d(t)_G, \mathcal{S}_n, K_F)$, which concludes the proof. □

A.3 Proof Proposition 3.2

Proof. Let $\phi_F(\mathbf{z}) = [1 + (\mathbf{z} - \widehat{\mathbf{m}}_0)\Sigma_F^{-1}(\mathbf{z} - \widehat{\mathbf{m}}_0)^T]^{-1}$. By Definition 3.18:

$$\begin{aligned} GKD(\tilde{x}(t), \mathcal{S}_n, K_F) &= \phi_F(\mathbf{z})\phi_F(\widehat{\mathbf{m}}_0) \\ &= \left([1 + (\mathbf{z} - \widehat{\mathbf{m}}_0)\Sigma_F^{-1}(\mathbf{z} - \widehat{\mathbf{m}}_0)^T]^{-1}\right)\left([1 + (\widehat{\mathbf{m}}_\mathbf{0} - \widehat{\mathbf{m}}_0)\Sigma_F^{-1}(\widehat{\mathbf{m}}_0 - \mu_F)^T]^{-1}\right) \\ &= [1 + (\mathbf{z} - \widehat{\mathbf{m}}_0)\Sigma_F^{-1}(\mathbf{z} - \widehat{\mathbf{m}}_0)^T]^{-1}, \end{aligned}$$

where $\widehat{\mathbf{m}}_0 = \mu_F$ the estimated center of the distribution F as defined in Def. 3.2. The Mahalanobis depth satisfy the property of *monotonicity relative to the deepest point* –see Zuo and Serfling (2000) for further details–, which implies that $M_hD(\mathbf{x}, F) = [1 + (\mathbf{x} - \mu_F)\Sigma_F^{-1}(\mathbf{x} - \mu_F)]^{-1}$ is an asymptotic f-monotone function –see Definition 3.16– what concludes the proof. □

A.4 Proof Proposition 3.3

Proof. $GKD(\tilde{x}(t), \mathcal{S}_n, K_F) = \phi_F(\mathbf{z})\phi_F(\widehat{\mathbf{m}}_0)$

Consider $\tilde{x}(t)$ be a sample curve and $\mathbf{z}$ the its correspnding $\mathcal{H}_d$ representation, and the sample version of the h-MD is given by $h - MD(\tilde{x}(t), \mathcal{S}_n) = \frac{1}{n}\sum_{l=1}^{n} K_h(\|\tilde{x}(t) - \tilde{x}_l(t)\|)$.

Let $\phi_F(\mathbf{z}) = \frac{1}{n}\sum_{l=1}^{n} K_h(\|\tilde{x}(t) - \tilde{x}_l(t)\|)$; then by definition of the GKD

$$GKD(\tilde{x}(t), \mathcal{S}_n, K_F) = \phi_F(\mathbf{z})\phi_F(\mathbf{m}_0),$$
$$= \left(\frac{1}{n}\sum_{l=1}^{n} K_h(\|\tilde{x}(t) - \tilde{x}_l(t)\|)\right)\left(\frac{1}{n}\sum_{l=1}^{n} K_h(\|\tilde{m}_0(t) - \tilde{x}_l(t)\|)\right),$$

where $\tilde{m}_0(t)$ is the modal curve asociated to $\mathbf{m}_0 \in \mathcal{H}_d$. Therefore $\tilde{m}_0(t)$ satisfies that:

$$\tilde{m}_0(t) = arg\max_{\tilde{x}(t)} h - MD(\tilde{x}(t), \mathcal{S}_n)$$

Using the normalized version of the h-mode depth –proposed by the authors in Cuevas et al. (2007):

$$\frac{h - MD(\tilde{x}(t), \mathcal{S}_n)) - min(h - MD(\tilde{x}(t), \mathcal{S}_n))}{max(h - MD(\tilde{x}(t), \mathcal{S}_n)) - min(h - MD(\tilde{x}(t), \mathcal{S}_n))},$$

it holds that: $h - MD(\tilde{x}(t), \mathcal{S}_n) = 1$. Therefore,

$$GKD(\tilde{x}(t), \mathcal{S}_n, K_F) = \frac{1}{n}\sum_{l=1}^{n} K_h(\|\tilde{x}(t) - \tilde{x}_l(t)\|),$$

which implies that the $h - MD$ is a particular case of the GKD. It remains to be prooved that the h-mode Depth is a density Kernel and therefore satisfy Definition 3.16. As well defined depth function, the h-mode Depth satisfy the property of *monotonicity relative to the deepest point* –see Nieto-Reyes and Battey (2016)–, which implies that $h - MD(\tilde{x}(t), \mathcal{S}_n) = \sum_{l=1}^{n} K_h(\|\tilde{x}(t) - \tilde{x}_l(t)\|)$ is an asympototic f-monotone function –see Definition 3.16– what concludes the proof. □

Appendix B

Appendix to Chapter 4

B.1 Proof Theorem 4.1

Proof Theorem 1. Consider the following optimization problem:

$$\min_{\beta_1,\ldots,\beta_n} \sum_{i=1}^{n} \beta_i \widehat{h}_\alpha(\Delta_{\mathbf{z}_i}) \quad \text{s.t.} \quad \sum_{i=1}^{n} \beta_i = n(1-\nu) \text{ and } 0 \leq \beta_i \leq 1 \text{ for } i = 1, \ldots, n. \tag{B.1}$$

For the sake of simplicity, consider first the case where $n(1-\nu) \in \mathbb{N}$. Let q^* be the $1-\nu$ quantile of the S_n sample. Then, it can be shown that $\beta_i^* = 1$ if $\hat{h}_\alpha(\Delta_{\mathbf{z}_i}) \leq q^*$ and $\beta_i^* = 0$ if $\hat{h}_\alpha(\Delta_{\mathbf{z}_i}) > q^*$ is a solution for the problem stated in Equation B.1. As a consequence

$$\frac{1}{n}\sum_{i=1}^{n} I(\mathbf{z}_i) = \frac{1}{n}\sum_{i=1}^{n} \beta_i^*.$$

From the constraint in Equation B.1 it holds that $\sum_{i=1}^{n} \beta_i^* = n(1-\nu)$, and then

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \beta_i^* = \lim_{n\to\infty} \frac{1}{n} n(1-\nu) = 1-\nu$$

For the case $n(1-\nu) \notin \mathbb{N}$, it holds that $\begin{cases} \beta_i = 1, & \text{if } \hat{h}_\alpha(\Delta_{\mathbf{z}_i}) < q^* \\ \beta_i = n(1-\nu) - [n(1-\nu)], & \text{if } \hat{h}_\alpha(\Delta_{\mathbf{z}}) = q^* \\ \beta_i = 0, & \text{if } \hat{h}_\alpha(\Delta_{\mathbf{z}_i}) > q^* \end{cases}$

where $[z]$ stands for the largest integer not greather than x. Therefore, the number

of β_i^*'s equating to 1 is $[n(1-\nu)]$ and

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} I(\mathbf{z}_i) = \lim_{n\to\infty}\frac{1}{n}([n(1-\nu)]\times 1+1) = \lim_{n\to\infty}\frac{[n(1-\nu)]}{n} = 1-\nu.$$

Finally we show that $\rho^* = q^*$. The dual problem of B.1 is:

$$\max_{b,\epsilon_1,\ldots,\epsilon_n} n(1-\nu)b - \sum_{i=1}^{n}\epsilon_i \quad \text{s.t. } \hat{h}_\alpha(\Delta_{\mathbf{z}_i}) \geq b-\epsilon_i,\ \epsilon_i \geq 0 \text{ for } i=1,\ldots,n. \tag{B.2}$$

By the fundamental theorem of duality, the objective functions of the problems stated in Equations B.1 and B.2 take the same value at their solutions and, as a consequence, $b^* = q^*$ (see Muñoz and Moguerza (2006)). Since problem B.2 difers from problem 4.2 just in the scaling of the objective function, it holds that $\rho^* = b^*$, which concludes the proof. □

Appendix C

Appendix to Chapter 5

In the numerical experiment we have considered the Gaussian Kernerl to estimate the autocorrlation operator Ψ_K. In that sense, we designed a Monte–Carlo experiment to define the optimal value for the Kernel parameter σ, of the kernel function: $K(x, y) = \exp^{\sigma(x-y)^2}$. For each simulated functional time series data set we conducted a search over a grid of the parameter σ. The results are presented in Figure C.1.

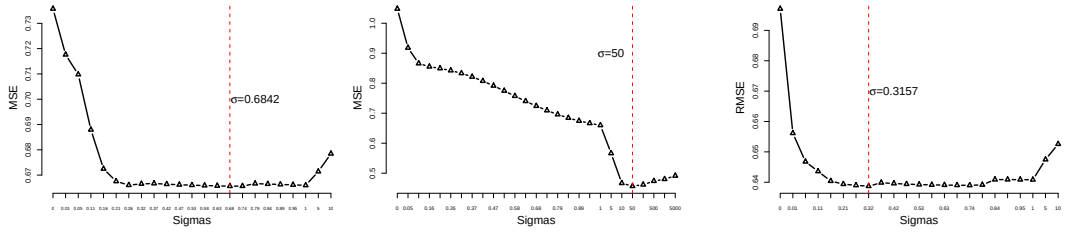


Figure C.1: Monte-Carlo Results: MSE for a grid of values for the kernel parameter $sigma$. FAR(1) process (left); AR–coefficients (middle); Wiener process (right).

Bibliography

Antoniadis, A., Paparoditis, E., and Sapatinas, T. (2006). A functional wavelet–kernel approach for time series prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5):837–857.

Arribas-Gil, A. and Romo, J. (2014). Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15(4):603–619.

Beirlant, J., Dudewicz, E. J., Györfi, L., and Van der Meulen, E. C. (1997). Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39.

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Berrendero, J. R., Cuevas, A., and Torrecilla, J. L. (2016). Variable selection in functional data classification: a maxima-hunting proposal. *Statistica Sinica*, pages 619–638.

Boos, D. D. (2004). L-statistics. *Encyclopedia of Statistical Sciences*, 7.

Bosq, D. (2012). *Linear processes in function spaces: theory and applications*, volume 149. Springer Science & Business Media.

Bouveyron, C. and Jacques, J. (2011). Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, 5(4):281–300.

Cano, J., Moguerza, J. M., Psarakis, S., and Yannacopoulos, A. N. (2015). Using statistical shape theory for the monitoring of nonlinear profiles. *Applied Stochastic Models in Business and Industry*, 31(2):160–177.

Chakraborty, A. and Chaudhuri, P. (2014a). The deepest point for distributions in infinite dimensional spaces. *Statistical Methodology*, 20:27–39.

Chakraborty, A. and Chaudhuri, P. (2014b). On data depth in infinite dimensional spaces. *Annals of the Institute of Statistical Mathematics*, 66(2):303–324.

Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49.

Cuesta-Albertos, J. A. and Nieto-Reyes, A. (2008). The random tukey depth. *Computational Statistics & Data Analysis*, 52(11):4979–4988.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, 147:1–23.

Cuevas, A., Febrero, M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496.

David, H. A. and Nagaraja, H. N. (2004). Order statistics. *Encyclopedia of Statistical Sciences*, 9.

Degras, D. (2017). Simultaneous confidence bands for the mean of functional data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(3):e1397.

Didericksen, D., Kokoszka, P., and Zhang, X. (2012). Empirical properties of forecasts with the functional autoregressive model. *Computational statistics*, 27(2):285–298.

Efron, B. and Tibshirani, R. (1994). An introduction to the bootstrap (crc, fl).

Engle, R., . G. C. (1991). *Long-run economic relationships: Readings in cointegration*. Oxford University Press.

Febrero-Bande, M., de la Fuente, M. O., et al. (2012). Statistical computing in functional data analysis: The r package fda. usc. *Journal of statistical Software*, 51(4):1–28.

Febrero-Bande, M. and Oviedo de la Fuente, M. (2013). Package 'fda. usc': Functional data analysis and utilities for statistical computing. *R package version 1.4.0*.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.

Fraiman, R., Gimenez, Y., and Svarc, M. (2016). Feature selection for functional data. *Journal of Multivariate Analysis*, 146:191–208.

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test*, 10(2):419–440.

Fresoli, D., Ruiz, E., and Pascual, L. (2014). Bootstrap multi-step forecasts of non-gaussian var models. *International Journal of Forecasting*.

Hall, P. and Heyde, C. C. (2014). *Martingale limit theory and its application*. Academic press.

Hall, P. and Hooker, G. (2016). Truncated linear models for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):637–653.

Härdle, W., Horowitz, J., and Kreiss, J.-P. (2003). Bootstrap methods for time series. *International Statistical Review*, 71(2):435–459.

Hero, A. O. (2007). Geometric entropy minimization (gem) for anomaly detection and localization. In *Advances in Neural Information Processing Systems*, pages 585–592.

Hörmann, S. and Kokoszka, P. (2012). Functional time series. In *Handbook of statistics*, volume 30, pages 157–186. Elsevier.

Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*, volume 200. Springer Science & Business Media.

Hsing, T. and Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons.

Hu, Y., Wang, Y., Wu, Y., Li, Q., and Hou, C. (2011). Generalized mahalanobis depth in the reproducing kernel hilbert space. *Statistical Papers*, 52(3):511–522.

Hyndman, R. (2017). Demography package. *R Foundation for Statistical Computing: Vienna, Austria*.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, 50(2):120–126.

Hyndman, R. J. and Shang, H. L. (2009). Forecasting functional time series. *Journal of the Korean Statistical Society*, 38(3):199–211.

Ieva, F. and Paganoni, A. M. (2013). Depth measures for multivariate functional data. *Communications in Statistics-Theory and Methods*, 42(7):1265–1276.

J Mercer, B. (1909). Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Phil. Trans. R. Soc. Lond. A*, 209(441-458):415–446.

Jacques, J. and Preda, C. (2013). Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing*, 112:164–171.

Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255.

Kargin, V. and Onatski, A. (2008). Curve forecasting by functional autoregression. *Journal of Multivariate Analysis*, 99(10):2508–2526.

Karhunen, K. (1946). Zur spektraltheorie stochastischer prozesse. *Ann. Acad. Sci. Fennicae, AI*, 34.

Keogh, E., Wei, L., Xi, X., Lonardi, S., Shieh, J., and Sirowy, S. (2006). Intelligent icons: Integrating lite-weight data mining and visualization into gui operating systems. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 912–916. IEEE.

Kreiss, J.-P. and Lahiri, S. N. (2012). Bootstrap methods for time series. In *Handbook of statistics*, volume 30, pages 3–26. Elsevier.

Kreiss, J.-P. and Paparoditis, E. (2011). Bootstrap methods for dependent data: A review. *Journal of the Korean Statistical Society*, 40(4):357–378.

Li, J., Cuesta-Albertos, J. A., and Liu, R. Y. (2012). Dd-classifier: Nonparametric classification procedure based on dd-plot. *Journal of the American Statistical Association*, 107(498):737–753.

Liu, J., Zhong, L., Wickramasuriya, J., and Vasudevan, V. (2009). uwave: Accelerometer-based personalized gesture recognition and its applications. *Pervasive and Mobile Computing*, 5(6):657–675.

Liu, R. Y. et al. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414.

Liu, R. Y., Parelius, J. M., Singh, K., et al. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference,(with discussion and a rejoinder by liu and singh). *The annals of statistics*, 27(3):783–858.

Loève, M. (1946). Fonctions aléatoires à décomposition orthogonale exponentielle. *La Revue Scientifique*, 84:159–162.

López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734.

López-Pintado, S. and Romo, J. (2011). A half-region depth for functional data. *Computational Statistics & Data Analysis*, 55(4):1679–1695.

Lopez-Pintado, S. and Torrente, A. (2013). depthtools: Depth tools package. *R package version 0.4. URL: http://CRAN. R-project. org/package= depthTools*.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55.

Maronna, R. A., Martin, R. D., Yohai, V. J., and Salibián-Barrera, M. (2018). *Robust statistics: theory and methods (with R)*. Wiley.

Martos, G. and Hernández, N. (2018). bigdatadist: Distances for machine learning and statistics in the context of big data. *R package version 1.0.*

Martos, G., Muñoz, A., and González, J. (2014). Generalizing the mahalanobis distance via density kernels. *Intelligent Data Analysis*, 18(6S):S19–S31.

Moguerza, J. M. and Muñoz, A. (2006). Support vector machines with applications. *Statistical Science*, pages 322–336.

Muñoz, A. and González, J. (2010). Representing functional data using support vector machines. *Pattern Recognition Letters*, 31(6):511–516.

Muñoz, A., Hernández, N., Moguerza, J., and Martos, G. (2018). Combining entropy measures for anomaly detection. *Entropy*, 20(9):698.

Muñoz, A. and Moguerza, J. (2006). Estimation of high-density regions using one-class neighbor machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):476–480.

Nagbe, K., Cugliari, J., and Jacques, J. (2018). Short-term electricity demand forecasting using a functional state space model. *Energies*, 11(5):1120.

Nagy, S. (2015). Consistency of h-mode depth. *Journal of Statistical Planning and Inference*, 165:91 – 103.

Nieto-Reyes, A. and Battey, H. (2016). A topologically valid definition of depth for functional data. *Statistical Science*, 31(1):61–79.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1(6):327–332.

Olszewski, R. T. (2001). Generalized feature extraction for structural pattern recognition in time-series data. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE.

Paparoditis, E. et al. (2018). Sieve bootstrap for functional time series. *The Annals of Statistics*, 46(6B):3510–3538.

Pascual, L., Romo, J., and Ruiz, E. (2004). Bootstrap predictive inference for arima processes. *Journal of Time Series Analysis*, 25(4):449–465.

Pini, A. and Vantini, S. (2017). Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, 29(2):407–424.

Ramsay, J. (1982). When the data are functions. *Psychometrika*, 47(4):379–396.

Ramsay, J. O. (2006). *Functional data analysis*. Wiley Online Library.

Rényi, A. et al. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Segaert, P. (2018). mrfdepth: Depth measures in multivariate, regression and functional settings. *R package version 1.0.7*.

Serfling, R. (2002). A depth function and a scale curve based on spatial quantiles. *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pages 25–38.

Serfling, R. (2006). Depth functions in nonparametric multivariate inference. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 72:1.

Sguera, C., Galeano, P., and Lillo, R. (2014). Spatial depth-based classification for functional data. *Test*, 23(4):725–750.

Shang, H. and Hyndman, R. (2013). Fds: functional data sets. r package version 1.7.

Shang, H. L. (2018). Bootstrap methods for stationary functional time series. *Statistics and Computing*, 28(1):1–10.

Stine, R. A. (1987). Estimating properties of autoregressive forecasts. *Journal of the American statistical association*, 82(400):1072–1078.

Tan, C. W., Webb, G. I., and Petitjean, F. (2017). Indexing and classifying gigabytes of time series under time warping. In *Proceedings of the 2017 SIAM international conference on data mining*, pages 282–290. SIAM.

Tarabelloni, N. (2018). roahd: Robust analysis of high dimensional data. *R package version 1.4*.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the international congress of mathematicians*, volume 2, pages 523–531.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*. Chapman and Hall/CRC.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295.

Wolf, M. and Wunderli, D. (2015). Bootstrap joint prediction regions. *Journal of Time Series Analysis*, 36(3):352–376.

Xie, T., Narabadi, N., Hero, A. O., et al. (2016). Robust training on approximated minimal-entropy set. *arXiv preprint arXiv:1610.06806*.

Zhang, J., Olive, D. J., and Ye, P. (2012). Robust covariance matrix estimation with canonical correlation analysis. *International Journal of Statistics and Probability*, 1(2):119.

Zhang, J.-T., Chen, J., et al. (2007). Statistical inferences for functional data. *The Annals of Statistics*, 35(3):1052–1079.

Zuo, Y. et al. (2003). Projection-based depth functions and associated medians. *The Annals of Statistics*, 31(5):1460–1490.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Annals of statistics*, pages 461–482.